Dynamic Coordination of the New York City Taxi to Optimize the Revenue of the Taxi Service

Jacky P.K. Li
Vrije Universiteit Amsterdam,
Faculty of Science,
Amsterdam, The Netherlands
Email: jacky.li@vu.nl

Sandjai Bhulai
Vrije Universiteit Amsterdam,
Faculty of Science,
Amsterdam, The Netherlands
Email: s.bhulai@vu.nl

Theresia van Essen
Delft Institute of Applied Mathematics,
Delft, The Netherlands
Email: J.T.vanEssen@tudelft.nl

Abstract—**Taxis are an essential component of the transportation system in most urban centers. The ability to optimize the efficiency of routing represents an opportunity to increase revenue for taxi drivers. Vacant taxis on the road waste fuel, represent uncompensated time for the taxi driver and create unnecessary carbon emissions while also generating additional traffic in the city. In this paper, we utilize Markov Decision Processes to optimize the revenue of taxi drivers through better routing. We present a case study utilizing real-world New York City Taxi data with several experimental evaluations of our model. We achieve approximately 10% improvement in efficiency using data from the month of January, representing the best scenario for an arbitrary taxi driver in that particular period of time. These results also provide a better understanding of how optimization strategies may differ during different times of the day. In the second half of the paper, we present a dynamic fleet management model that can handle random load arrivals with multiple vehicles in Manhattan in a period of 30 minutes. The fleet management problem decomposes into a sequence of time-indexed min-cost network flow subproblems that naturally yield integer solutions. These two methods may have important implications in the field of self-driving vehicles.**

Keywords–New York taxi service; revenue optimization; optimal routing; Markov decision processes; linear programming; min-cost network flow problem.

I. Introduction

In New York City, there are over 485,000 passengers taking taxis per day, equating to over 175 million trips per year [1], [2]. Creating an efficient way to transport passengers through the city is of utmost importance. Taxi drivers cannot control a passenger's destination but can make better decisions using optimal routing. This consequently leads to a reduction in costs and carbon emissions.

Previous studies have focused on developing recommendation systems for taxi drivers [3]–[8]. Several studies use the GPS system to create recommendations for both the drivers and the passengers to increase profit margins and reduce seek times [5], [7]–[9]. Ge et al. [10] and Ziebart et al. [11] gather a variety of information to generate a behavior model to improve driving predictions. Ge et al. [3] and Tseng et al. [12] measure the energy consumption before finding the next passenger. Castro et al. [9], Altshuler et al. [13], Chawla et al. [14], Huang et al. [15], and Qian et al. [16] learn knowledge from taxi data for other types of recommendation scenarios such as fast routing, ride-sharing, or fair recommendations.

Most of the papers above focus on optimizing measures for the immediate next trip. Rong et al. [4] investigate how to inform business strategies from the historical data to increase revenues of the taxi drivers using Markov decision processes (MDPs). Their research model uses historical data to estimate the probability of finding a passenger and its location for drop-off as the necessary parameters for the MDP model. For each one-hour time slot, the model learns a different set of parameters for the MDP from the data and finds the optimal move for the vacant taxi to maximize the total revenue in that time slot. At each state, the MDP model uses a combination of location, time, current and previous actions. The vacant taxi can travel to its neighboring locations and cruise through the grid to seek for the next passenger. Using dynamic programming to solve the MDP, the output of the model recommends the best actions for the taxi driver to take at each state.

Tseng et al. [12] examine the viability of electric taxis in New York City by using MDPs. Due to the radius limitation of electric taxis before each charge, they examine the profitability of replacing conventionally fueled taxis with electric taxis. The research model uses OpenStreetMap (OSM) to assign each pick-up and drop-off into the nearest junctions. The advantage of using OSM is that it is able to identify the number of available taxis at the junction without extra calculations. The research is concentrated on energy consumption; the actions become infeasible if the electric vehicle runs out of energy.

Analysis of real taxi data shows that there are significant differences in demand between certain periods of the day. The aforementioned research has not taken the effect of this demand variation into account. The contribution of our model is that we extend the research by Rong et al. [4] in this direction. We analyze the New York City Taxi data and study the differences in optimal policy and revenue for the demand between weekdays, weekends, day shifts, and night shifts. From these observations, we can infer relevant policies for taxi drivers based on the shift that they work in.

In addition to using Markov Decision Process on the New York City data, in the second half of this paper, we introduce a dynamic fleet management model to solve the vehicle coverage problem. The contribution of our model is that we extend the research by Topaloglu et al. [17] in this direction. Dynamic resource allocation problems assign a set of resources to determine tasks over a period of time. Such problems arise

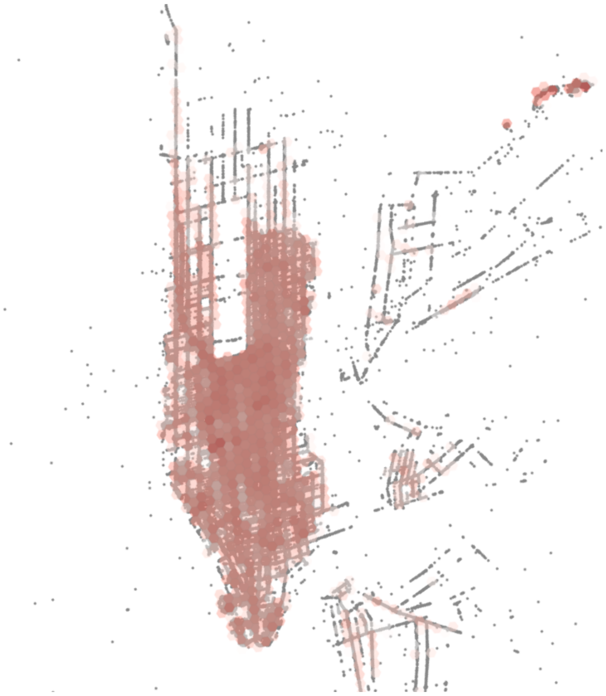

Figure 1. Rotated Manhattan with the total revenue for the NYC Taxi by pick-up location in January 2013.

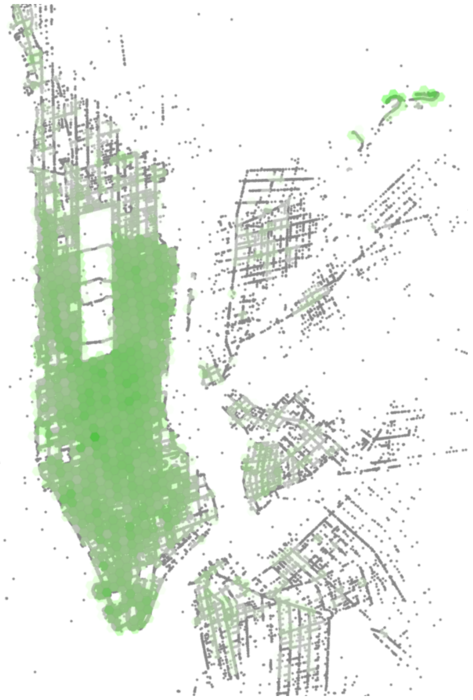

Figure 2. Rotated Manhattan with the total revenue for the NYC Taxi by drop-off location in January 2013.

in many fields such as dynamic fleet management [18], [19], [20], product distribution [21], machine scheduling [22], and personnel management [23]. In the second half of this paper, we are confronted with this problem within the context of managing NYC taxis to serve customers who request a ride. We assume the total business time is equal to the sum of the total occupancy time and the total seeking time. Fundamentally, if we can satisfy as many customer ride requests and minimize the seeking time, this would provide the maximum profit in the overall system.

The deterministic version of this problem is the min-cost integer problem. The linear and integer versions for the min-cost “multi-commodity-flow” problem have been studied extensively in [19] and [24].

The paper is structured as follows. In Section II, we do data analysis on the New York Taxi dataset. This provides input for our MDP, which is explained in Section III. We assess the performance of the MDP in Section IV, where we conduct numerical experiments. In section V, we introduce dynamic resource allocation method to solve our min-cost integer problem, and we assess the performance of the linear program. Finally, the paper is concluded in Section VI and the future discussion in Section VII.

II. Dataset

In our research, we selected to use New York City Taxi data in 2013 provided by NYC Taxi & Limousine Commission [2], which includes the encrypted taxi ID, encrypted medallion and the exact GPS location. Due to privacy issues, the taxi ID and medallion were omitted from the data since 2013. In order to compare our model to each individual taxi, the taxi ID and medallion were important. This is one of the reasons we decided to use 2013 data.

From the data, we use 14,776,615 taxi rides collected in New York City over a period of one month (January 2013) [2]. For illustrative purposes, we pick the month of January in this paper, however, the model allows any month to be used as input. From each ride record, we use the following fields: taxi ID, pick-up time, pick-up longitude, pick-up latitude, drop-off time, drop-off longitude, drop-off latitude, the number of passengers per ride, average velocity, trip distance, traveling time, and fare amount. We omit the records containing missing or erroneous GPS coordinates. Records that represent rides that started or ended outside Manhattan, as well as trip durations longer than 1 hour and trip distances greater than 100 kilometers are omitted as well. Furthermore, we collect the drivers who drive for six to nine hours consistently to yield a clean dataset containing approximately 13.5 millions taxi rides. We observe that most of the pick-up locations are in the Manhattan area.

We concentrate on the island of Manhattan area in NY. This area imposes a rectangular grid of avenues and streets. However, the city’s avenues are not parallel to the true north and south. For that reason, we tilted the map by 28.899 degrees according to Petzold et al. [25]. This creates blocks with the same grid system in most areas. We discretize the grid into a 50×50 grid, making each block in the grid approximately 300 meters $\times$ 300 meters. The choice for a block size of 300 meters is based on the assumption that a taxi can traverse this distance within 1 minute. Figure 1 shows the total revenue for the taxis by the pick-up location with the rotated map. Figure 2 indicates the total revenues of the drop-off location, and it shows that Lower Manhattan, along with the airport are the largest revenue generators and the drop-off location has spread to the mid-Manhattan area and also Brooklyn area.

The state of a taxi can be described by two parameters:

Table I. Revenue Efficiency E_{rev} (in \$/minute).

	Weekday dayshift	Weekday nightshift	Weekend dayshift	Weekend nightshift	Overall
Top 10%	0.59203	0.62408	0.60111	0.64646	0.60869
Mean	0.49985	0.52232	0.50252	0.54871	0.50565
Standard Deviation	0.07253	0.08011	0.07787	0.07799	0.08088
Bottom 10%	0.41028	0.42174	0.40426	0.44978	0.40572

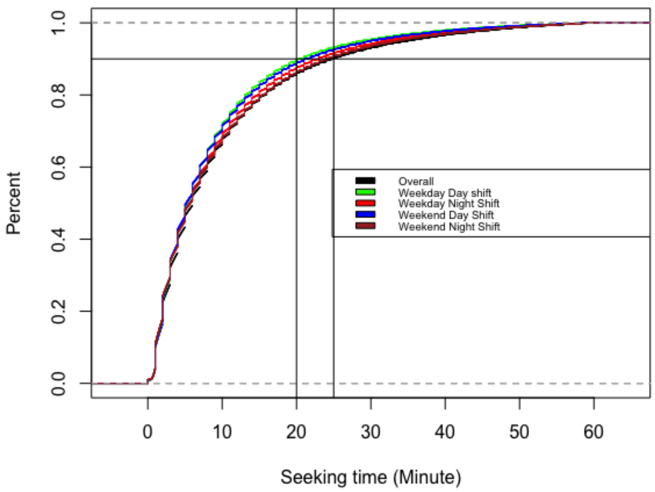

Figure 3. Seeking time for the models.

the current location, which is an element of the set $L = \{(1,1),\ldots,(50,50)\}$ grid and the current time, which comes from the set $T = \{1,\ldots,60\}$. We will denote the system state in our MDP model as $s = (x, y, t)$, which we will elaborate on in Section IV and in Section V.

III. METHODOLOGY FOR MARKOV DECISION PROCESS

A. Performance indicators

In this section, we present performance indicators of the taxi drivers. This will be used in the MDP to optimize the routing decision of each taxi driver. Hence, the performance indicators will be dependent on the routing policy that is being applied by the taxi drivers. To improve readability, we drop the dependency on the policy in the notation and use it only in cases where it benefits clarity.

We calculate the total business time of each taxi driver per shift. The total business time (denoted as T_{bus}) is equal to the sum of the total occupancy time (T_{occupy}) and the total seeking time (T_{seek}):

$$T_{\text{bus}} = T_{\text{occupy}} + T_{\text{seek}}. \tag{1}$$

The total occupancy time T_{occupy} is the sum of all the trip durations with passengers of a taxi per day. And the total seeking time T_{seek} is the time between each trip. Figure 3 depicts the overall T_{seek} and the graphs in which we distinguish between the weekday, weekend, day shift, and the night shift. Based on the data, we assume 90% of the seeking times are shorter than 20 minutes for the day shift and shorter than 25 minutes for the night shift. Therefore, we discount any seeking time that is over 30 minutes as we assume those are the breaks for the drivers.

Logically, the T_{bus} is approximately the same for each taxi driver. To increase the revenue, the taxi drivers aim to have the maximal T_{occupy} and the minimal T_{seek}. We define the revenue efficiency E_{rev} metric as the revenue earned divided by the total taxi driver's business time. This is expressed as follows:

$$E_{\text{rev}} = \frac{M}{T_{\text{bus}}} = \frac{M}{T_{\text{occupy}} + T_{\text{seek}}}, \tag{2}$$

where M denotes the total money earned by the taxi driver during that period.

To illustrate the consistency of the taxi driver, we concentrate on the drivers who work between six hours to nine hours during the month of January. From that data, we generate the data of P_{find}, P_{dest}, T_{drive}, r (parameters of our MDP to be described in the next section) of each model and identify the top 10% and bottom 10% drivers in each model.

Table I indicates the revenue efficiency of the top 10% and bottom 10% distinguished by weekday, weekend, day shift, night shift, and the overall efficiency. Based on the table, there is an approximate 20% difference between the performance of the top 10% and bottom 10% drivers. The previous studies that were mentioned above (see, e.g., [5], [9], [12], [14], [15]) attribute the difference between the performance by the top and bottom 10% of drivers to the seeking time of the taxi drivers. This warrants research to determine if our model can provide a better solution for the taxi drivers for seeking passengers.

IV. MATHEMATICAL MODEL FOR THE MARKOV DECISION PROCESS

In order to model the taxi service in New York City, we adopt the framework of MDPs. This framework allows us to deal with the uncertain demand over the different periods in the grid, and to model them explicitly. The MDP is a stochastic decision process with a set S of states and a set A of possible actions that transition the states from one to another. Each action will correspond to the process of the current state to the new state with a probability transition function and a reward function. The collection of optimal actions for each state is called the policy, which maximizes the total reward over several numbers of steps. The objective of our model is to minimize the seeking time for the taxi to maximize the expected revenues.

A. System States

The state for a taxi is described by its current location and the current time. The details are explained as follows.

Location $(x, y) \in L = \{1,\ldots,50\} \times \{1,\ldots,50\}$: the area is divided into a grid of 50×50 grid cells;

Time $t \in T = \{1, \ldots, 60\}$: we use minutes as the interval of a time slot, and a total of 1 hour as time horizon.

Each pick-up and drop-off location is assigned to a grid cell. We remove the records that contain 1) incomplete data information, 2) trip distance over 100 kilometers, 3) trip durations over 60 minutes, 4) pick-up and drop-off locations with the same coordinates, 5) pick-up and drop-off locations outside the grid, and 6) shifts that are shorter than six hours and longer than nine hours.

We denote the system state of our MDP model as $s = (x, y, t)$, and the collection of all admissible states is denoted by S.

B. Actions

The admissible actions from a given state s have nine possibilities to choose from. We use numbers $1, \ldots, 9$ to index the directions. The actions are mapped to directions in which the taxi moves as follows:

7	8	9
4	5	6
1	2	3

,

where, e.g., action 9 moves the taxi to the neighboring north-east location and action 5 is the current location of the taxi.

C. Parameters of the MDP model

In this subsection, we state the parameters used in the rest of MDP model.

The probability parameters are defined as:

- $P_{\text{find}}(x, y)$ describes the probability of successfully picking up a passenger in grid cell (x, y). We can calculate the probability of picking up a passenger in the cell by dividing the number of successful pick-ups in the cell $n_{\text{find}}(x, y)$ by the total number of times this cell is visited by a vacant taxi. The vacant taxi includes the taxis that drop off passengers in grid cell (x, y), denoted by $n_{\text{drop-off}}(x, y)$, and also the taxis that are seeking for passengers, denoted by $n_{\text{OSRM}}(x, y)$.
  To locate the vacant taxi every minute during the seeking trip, we use the API provided by Open Source Routing Machine [26], to estimate the coordinates. We use one-hour time slots between 12:00 to 13:00 for the day shift model and 0:00 to 1:00 for the night shift model. In our overall model, we took the average of the day time and night time models to estimate the number of vacant taxis at each grid during the month of January in 2013. Thus,
  $$P_{\text{find}} = \frac{n_{\text{find}}(x, y)}{n_{\text{find}}(x, y) + n_{\text{drop-off}}(x, y) + n_{\text{OSRM}}(x, y)}.$$
- $P_{\text{dest}}(x, y, x', y')$ describes the probability of a passenger traveling from grid cell (x, y) to the grid cell (x', y'). To estimate the destination probability for a time slot, we calculate the number of trips between each pair of source and destination locations in that time slot and get a 50×50 matrix. The value is divided by the sum of the entire number of trips of the grid cells. Therefore, P_{dest} has the empirical probability distribution of a passenger choosing destination location (x', y') when he is picked up at location (x, y).

The time parameters are defined as:

- $T_{\text{seek}}(a)$: The required time to travel from one location to a neighboring location based on action $a \in A$. We assume that the average speed of seeking trips is approximately 300 meters per minute. Thus, a taxi can traverse on cell when $a = 2, 4, 5, 6, 8$, and hence $T_{\text{seek}}(a) = 1$ in this case. In case $a = 1, 3, 7, 9$, then we set $T_{\text{seek}}(a)$ equal to 2, due to the diagonal movement.
- $T_{\text{drive}}(x, y, x', y')$: The driving time from (x, y) to (x', y'). We can calculate the total driving time from grid cell (x, y) to grid cell (x', y') and then divide by the number of trips from grid cell (x, y) to grid cell (x', y'). We calculate T_{drive} individually for all models. From the calculation, there is approximately +15.67% driving time difference between the day shift model and the night shift model, and there is a +4.14% difference between the weekend and the weekday.
- We assume there is no waiting time for passengers to get in and out of the vehicle.

The reward is defined as:

- $r(x, y, x', y')$: The expected reward from grid cell (x, y) to grid cell (x', y'). Similar to T_{drive}, we calculate the average fare of the number of trips between each pair of source and destinations as the expected fare. Note that due to this definition, the reward does not depend on the action of the taxi driver. We calculate r separately for all models. Similarly to T_{drive}, there is approximately a +6.21% reward difference between the day shift model and the night shift model, and there is a +1.21% difference between the weekend and the weekday.

D. State transition function of the MDP model

The state transition function describes the probability that one moves from state (x, y, t) after taking decision a moves to state (x', y', t'). Assuming the current state is $s = (x, y, t)$ and action a is taken, there are two possible outcomes of the transition:

1) The taxi successfully finds a passenger in grid (x, y) within $T_{\text{seek}}(a)$ minutes. The taxi with the passenger goes to destination (x', y') with probability $P_{\text{dest}}(x, y, x', y')$. The taxi arrives at location (x', y') with $T_{\text{drive}}(x, y, x', y')$ as the total time used to travel from (x, y) to (x', y'). The taxi driver receives $r(x, y, x', y')$ as the expected reward. Then the taxi will start seeking for a passenger from grid cell (x', y'). In this case, the new state becomes $s' = (x', y', t + T_{\text{seek}}(a) + T_{\text{drive}}(x, y, x'y'))$.
2) The taxi does not find a passenger after $T_{\text{seek}}(a)$ minutes being in grid (x, y) with probability $1 - P_{\text{find}}(x, y)$. The taxi driver does not receive a reward and saves the driving time T_{drive}. The taxi driver starts to make the next action at grid cell (x', y'). Hence, the state of the taxi driver becomes $s' = (x', y', t + T_{\text{seek}}(a))$.

E. The objective function of the MDP model

The objective function of the MDP model is to maximize the total expected rewards starting from an initial state. The terminal states are the states with $t = 60$. No more actions can be taken once the system reaches the terminal states. The maximal expected reward for an action a in state $s = (x, y, t)$ is expressed as $V(s, a)$ shown in (3).

$$\begin{aligned} V(s,a) = {} & (1 - P_{\text{find}}(x,y)) \times \\ & \max_{a' \in A} V(x, y, t + T_{\text{seek}}(a), a') + \\ & \sum_{(x',y') \in L} P_{\text{find}}(x,y) \times P_{\text{dest}}(x,y,x',y') \times \\ & \big[r(x,y,x',y') + \\ & \max_{a' \in A} V(x', y', t + T_{\text{seek}}(a) + T_{\text{drive}}(x,y,x',y'), a') \big]. \end{aligned} \tag{3}$$

The optimal policy π^* is defined as:

$$\pi^*(s) = \arg\max \{V(s,a)\}, \tag{4}$$

and the optimal value function is given by

$$V^*(s) = V(s, \pi^*(s)). \tag{5}$$

F. Markov Decision Process Solution

In order to solve the Markov decision problem to derive the optimal policy, we employ dynamic programming to maximize the expected rewards. The algorithm starts from time $t = 60$ and then traces backward to time $t = 1$. The algorithm is listed in Algorithm 1.

Algorithm 1 Solving MDP using Dynamic Programming

Input: $L, A, T, P_{\text{find}}, P_{\text{dest}}, r, T_{\text{drive}}, T_{\text{seek}}$
Output: The best policy π^*

```
V is a $|L| \times |T|$ matrix; $V \leftarrow 0$
for $t = |T|$ to 1 do
        for all $(x, y) \in L$ do ▷ $s = (x, y, t)$
                $a_{max} \leftarrow a$ that maximizes $V(s, a)$
$\pi^*(s) \leftarrow a_{max}$
$V^*(s) \leftarrow V(s, a_{max})$
return $\pi^*$
```

G. Case study

In this section, we present our case study on the New York Taxi dataset. We evaluate the MDP for the expected reward based on the dataset from January 2013. We assume that the NYC taxis have two shifts per day and each shift is a 12-hour period. We analyze the taxi's expected reward in 1) the day-time shift within six to nine hours of its operating time, 5 am to 5 pm and 2) the night-time shift, 5 pm to 5 am and 3) the weekdays from Monday to Friday, and 4) the weekend from Friday to Sunday. After filtering the data, we have approximately 170,000, 205,000, 145,000, and 193,000 shifts, respectively, for the Weekday day-time shift, Weekday night-time shift, Weekend day-time shift, and Weekend night-time shift. Although the weekend has a fewer number of days in January, the total number of shifts of the weekend night time is almost the same as for the weekday night time. The optimal policy is depicted in Figure 4. The figure presents the optimal policy by the MDP model at the particular time and location.

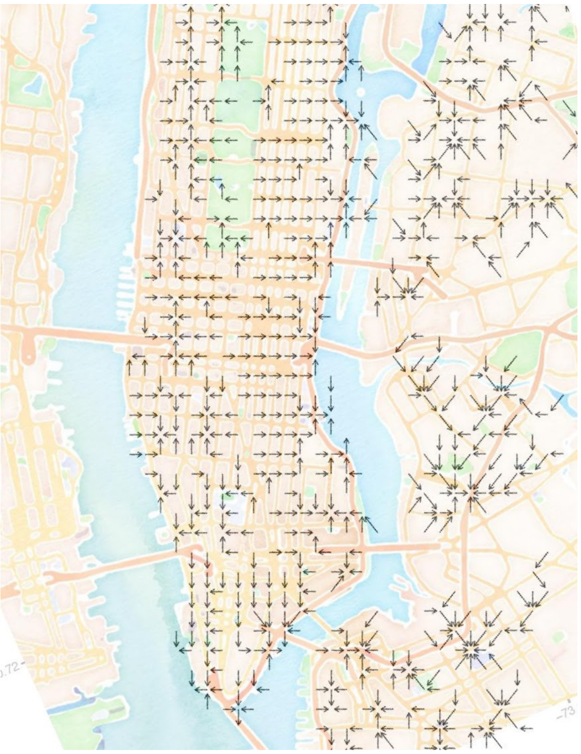

Figure 4. Recommended movements by the MDP model.

The results of the case study (see also Table II) shows that in our model

- $P_{\text{find}}(x, y)$ is 0.52267 which is 27.39% better than the bottom 10%, and it is 11.72% less effective than the top 10% for the Weekday day-time model.
- For the weekday night-time model, $P_{\text{find}}(x, y)$ is 0.50915 which is 20.73% better than the bottom 10%. It is 18.42% less effective than the top 10%.
- For the weekend day-time model, $P_{\text{find}}(x, y)$ is 0.51463 which is 27.30% better than the bottom 10%. It is 14.39% less effective than top 10%.
- For the weekend nighttime model, $P_{\text{find}}(x, y)$ is 0.45475 which is almost the same as the bottom 10% and it is 29.66% less effective than the top 10%.
- The overall model, $P_{\text{find}}(x, y)$ is 0.50030 which is 23.31% better than the bottom 10% and it is 17.81% less effective than top 10%.

The results of the case study show that our model is capable of reducing the time to find a passenger for a taxi driver significantly. Consequently, the end result is that the earnings of the taxi drivers increase. This benefit is expressed as approximately a 10% improvement in efficiency.

V. PART B: LINEAR PROGRAMMING

After using the MDP to optimize the revenue of a taxi service, the fundamental following question would be how

Table II. Revenue Efficiency E_{rev} (in \$/minute).

	Weekday	Weekday	Weekend	Weekend	Overall
	dayshift	nightshift	dayshift	nightshift	
Top 10%	0.59203	0.62408	0.60111	0.64646	0.60869
$P_{\text{find}}(x, y)$	0.52267	0.50915	0.51463	0.45475	0.50030
Bottom 10%	0.41028	0.42174	0.40426	0.44978	0.40572

many taxis are needed to satisfy all the demand? Part B of this paper will address this question.

The deterministic version of the taxi routing problem is by solving a max-profit integer "multi-commodity-flow" problem for each time period. The linear and integer versions of this problem have been studied extensively. This section empirically investigates the effectiveness of this approximation method when applied to resource allocation problems. These problems tend to get large easily with the number of possible states and resource types, and their multi-commodity nature presenting an unwelcome dimension of complexity.

We formulate our dynamic resource allocation problem using the language of Markov decision processes. We are modeling the taxi service in New York City with a fleet of taxis. At each decision epoch, a certain number of customers requests a ride (demand), each requesting to be taken from a certain location (x, y) to a destination (x', y'). For notational convenience, we denote (x, y) by i and denote $(x'y')$ by j. We assume the customers call in at the last minute, and very little information about the future requests is available in advance. We are required to serve every customer demand. However, if there are not enough vacant taxis within the same grid, the unsatisfied customer demands are not served. To handle this, we assume that the unsatisfied demands are lost, and we take the profit from serving a higher revenue demand to be the incremental profit from serving the demand with a taxi.

Our initial formulation assumes that all taxis take a single time period and all customers have the same taxi preferences. We also assume all the travel times take a single time period. For notational convenience, we assume that demand at a certain location can be served only by a taxi at the same location at the same time. For the rest of the section, we adopt the terminology that an empty taxi is "seeking".

A. Parameters of the Linear Programming Model

In this subsection, we state the parameters used in the rest of the model.

Location $(i, j) \in L = \{1, \ldots, 10\} \times \{1, \ldots, 10\}$: the area is divided into a grid of 10×10 grid cells; We implement a smaller grid compared to the first part of the paper in order to simplify the calculation process.

Time $t \in T = \{1, \ldots, 30\}$: we use minutes as the interval of a time slot, and a total of 30 minutes as time horizon.

- $D_{i,j,t}$ describes the number of **demand** that need to be carried from grid cell i to grid cell j at time period t.
- $S_{i,j,t}$ describe the number of **empty** taxis moving from grid cell i to grid cell j at time period t from the original dataset on January 15th, 2013.
- $x^{\mathrm{l}}_{i,j,t}$ describes the number of **loaded** taxis moving from grid cell i to grid cell j at time period t.
- $x^{\mathrm{e}}_{i,j,t}$ describes the number of **empty** taxis moving from grid cell i to grid cell j at time period t.
- $c^{\mathrm{l}}_{i,j}$ describes the net **reward** from an occupied taxi moving from grid cell i to grid cell j. We assume the profit is the same at any period of time t.
- $c^{\mathrm{e}}_{i,j}$ describes the **cost** of a vacant taxi moving empty from grid cell i to grid cell j. We assume the cost is the same at any period of time t. (Remark: In order to simplify the model, the **cost** is half of the **reward**.)
- $R_{i,j,t}$ describes the number of taxis in operation, including empty taxis and loaded taxis at time period t.

The deterministic version of the problem we are interested in can be written as:

$$\max \sum_{t \in T} \sum_{i,j \in L} (-c^{\mathrm{e}}_{i,j} x^{\mathrm{e}}_{i,j,t} + c^{\mathrm{l}}_{i,j} x^{\mathrm{l}}_{i,j,t}) \tag{6}$$

subject to

$$\begin{aligned}
&\sum_{j \in L} (x^{\mathrm{e}}_{i,j,1} + x^{\mathrm{l}}_{i,j,1}) = R_{i,1} && i \in L, \\
&-\sum_{j \in L} (x^{\mathrm{e}}_{j,i,t-1} + x^{\mathrm{l}}_{j,i,t-1}) + \sum_{j \in L} (x^{\mathrm{e}}_{i,j,t} + x^{\mathrm{l}}_{i,j,t}) = 0 && \\
& && i \in L, t \in \{2, \ldots, 30\}, \\
&x^{\mathrm{l}}_{i,j,1} \leq D_{i,j,t} && i, j \in L, t \in \{1, \ldots, 30\}, \\
&x^{\mathrm{e}}_{i,j,t}, x^{\mathrm{l}}_{i,j,t} \in \mathbb{Z}_+ && i, j \in L, t \in \{1, \ldots, 30\},
\end{aligned} \tag{7}$$

which is a special case of the min-cost integer multi-commodity flow problem.

B. Case Study 2

Similarly to the case study of the MDP model, we evaluate the linear programming approach based on the New York Taxi dataset of 2013. We concentrate on January 15th, 2013 from 12:00 pm to 12:30 pm. In our deterministic case study experiment, we formulate the problem as a max-profit integer problem (6). From the dataset, we generate the data of $D_{i,j,t}$ which is the number of demand from location i to location j at time t. We also generate the data of $S_{i,j,t}$ which is the number of empty taxi driving from location i to location j at time t to seek for the next passenger(s).

From Table III, the average of the demand is approximately 505.47 per minute, and the standard deviation is approximately 21.64 per minute within Manhattan. This indicates a consistent demand during this period. Due to all travel time, it lasted 1 minute in our model. Theoretically, we can assume approximately over 500 vehicles should satisfy all the odd number

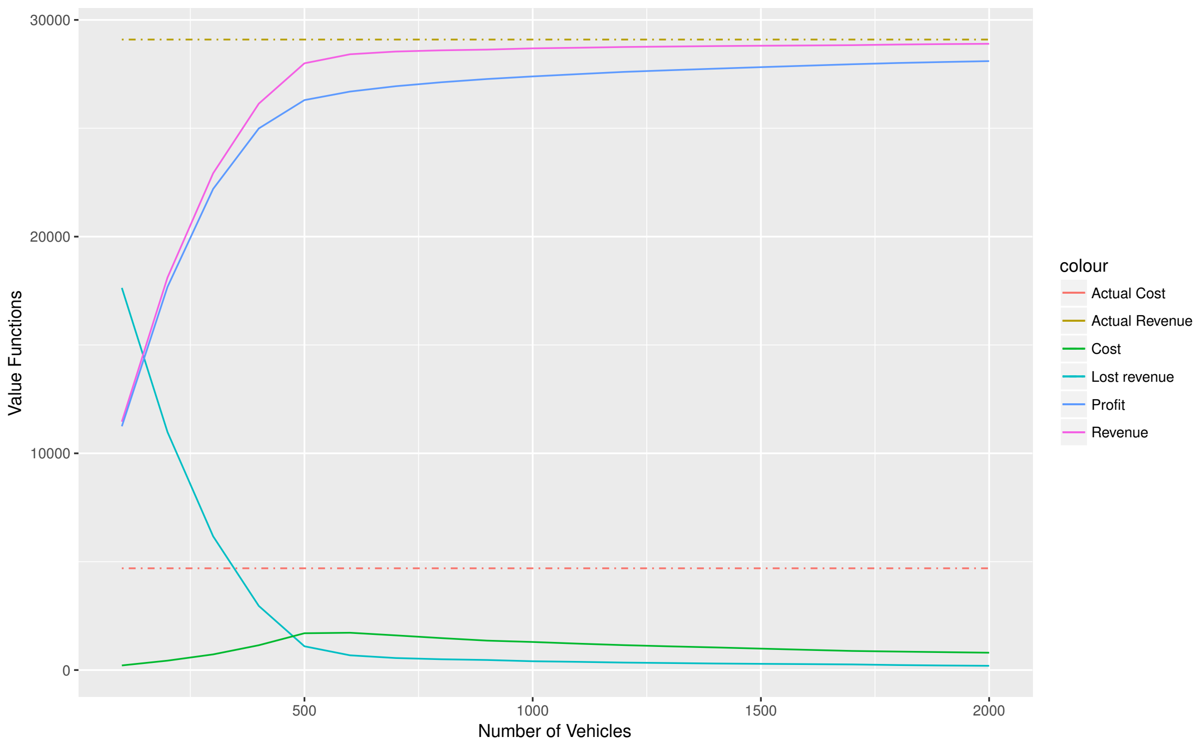

Figure 5. Overall model with 1 to 20 vehicles per grid in 30 minutes.

minute requests, and another 500 vehicles should satisfy the even number minute requests. To prove this theory, we ran our linear programming model by increasing the number of vehicles per grid at the initial minute at 12:00. We ran our model from 1 vehicle per grid to 30 vehicles per grid. In Figure 5, there is a clear indication of the difference between 100 vehicles to 1,200 vehicles. From this season, we further examine the actual revenue, revenue, actual cost, cost, profit, and lost revenue at each minute of the model with the number of vehicles from 300, 600, 900, and 1,200.

In order to provide a better understanding for our result, we calculate:

- Actual Revenue $= c^{\text{l}}_{i,j} \times D_{i,j,t}$
- Revenue $= c^{\text{l}}_{i,j} \times x^{*,\text{l}}_{i,j,t}$
- Actual Cost $= c^{\text{e}}_{i,j} \times S_{i,j,t}$
- Cost $= c^{\text{e}}_{i,j} \times x^{*,\text{e}}_{i,j,t}$
- Actual Profit $=$ Actual Revenue $-$ Actual Cost
- Profit = Revenue $-$ Cost
- Lost Revenue $= c^{\text{l}}_{i,j} \times [D_{i,j,t} - x^{*,\text{l}}_{i,j,t}]$,

Table III. Demand and Seeking from 12:00 pm to 12:30 pm on January 15th, 2013.

	Demand	Seeking
Minimum	468	371
Average	505.47	429.97
Standard Deviation	21.64	27.85
Maximum	540	485

where $x^{*,\text{e}}_{i,j,t}$ and $x^{*,\text{l}}_{i,j,t}$ are the optimal solutions for $x^{\text{e}}_{i,j,t}$ and $x^{\text{l}}_{i,j,t}$, respectively.

To set up the initial location of the vehicle, we spread the same number of vehicles in each grid, i.e., 300 vehicles indicate 3 vehicles in each grid over a 10×10 grid. Due to this initial condition, it will take a few minutes to relocate the vehicles properly in the grid. The results are clearly indicated in Figures 6, 7, 8, and 9. Figure 6 displays the total revenue of all vehicles per minute. The revenue of the 1,200 vehicles and 900 vehicles are similar to the actual revenue. With 300 vehicles in the grid, there are clearly not enough vehicles to satisfy all the demand. Surprisingly, 600 vehicles were able to receive similar revenue as the actual revenue. In Figure 7 the total cost of empty vehicles moving from i to j per minute is depicted. The dotted line shows the actual cost and our model indicates a clear lower cost than the actual cost. Figure 8 shows the revenue that is lost due to being unable to satisfy the demand. The 300 vehicles model is losing approximately 189.83 units per minute because it is unable to satisfy the demand. The rest of the model shows that the lost revenue is close to nothing. The most interesting observation is in Figure 9. There is a clear indication that the 1,200, 900, and 600 vehicles model do better than the actual profit. Thus, the vehicles move less overall to save on the cost and create bigger profit than the original data. The 300 vehicles model is the only model that makes less profit than the actual model.

VI. Conclusion

From the results of the case study in the MDP model, we observe that the weekend night time raises interesting discussion. It has a similar number of shifts as compared to the weekday night-time model, but the revenue efficiency did not improve compared to the bottom 10% drivers. A possible explanation might be that the experienced drivers would use their experience to look for the best location to seek customers. Consequently, the data may not have provided enough evidence to improve the bottom 10% drivers. In our data analysis, we found cases where there are pick-up and drop-off locations in the Hudson River. We can assume that this is an error in the GPS system. Similar to this issue, P_{dest} was estimated from a small number of trips from one location to another. This could sometimes result in a high probability, for instance, 1 of 3, would have created a 33% probability of going from one location to another. Further research is needed to develop methods to get a more accurate estimate.

In the second half of this paper, we use a linear programming to model the taxi service and determine the optimal policy to yield the best profit in the overall system. In Table IV, the demand column describes the total number of demand per minute and the seeking column describes the total number of vacant vehicles driving to seek for the next passenger. From Figure 6, we understand that the revenues are similar to 1,200 vehicles, 900 vehicles and 600 vehicles and it clearly shows that the 300 vehicles model is not sufficient to satisfy the demand and match the revenue of the actual data. To increase the profit, it requires to decrease the cost. From Table IV, the percentage difference of the profit is approximately $+14.20\%$ for the 1,200 vehicles model and is $+13.39\%$ for the 900 vehicles model, and is 12.30% for the 600 vehicles model. The percentage difference for the profit is -5.02% for the 300 vehicles model. Notice that the percentage difference was calculated without the first minute of the model, because the vehicles were distributed evenly in the model and it is not matching the demand of the locations during that first minute.

From Table IV, the lost revenue brings in some interesting observations. The lost revenue is defined as the revenue multiplied by the overall demand minus the optimal load, i.e., $= c^1_{i,j} \times [D_{i,j,t} - x^{*,1}_{i,j,t}]$. The average lost revenue is 1.48 units per minute for the 1,200 vehicles model, 3.07 units per minute for the 900 vehicles model, and 6.48 units per minute for the 600 vehicles model, and 189.83 units per minute for the 300 vehicles model. This clearly indicates that the 1,200 and 900 models creates a good result of not losing too many customers. In our conclusion, the 600 vehicles model, the 900 vehicles model, and the 1,200 vehicles model do not provide significant differences in terms of profit, revenue, and lost revenue. If we are focusing only on profit as our main priority, 600 vehicles would be sufficient enough to generate the profit that is similar to the 900 and 1,200 vehicles models. If we are focusing more toward the customer satisfaction, the 1,200 vehicles model would provide good profit and satisfy most of the customer requests during this 30-minute period.

VII. Future Discussion

As for future discussion, the demand that we harvested from the data is the demand that was satisfied that particular minute of January 15th, 2013 between 12:00 pm to 12:30 pm. This is the only demand that was satisfied by the yellow taxi. We could include all the demand that was satisfied by Uber or other vehicle services to see if all the yellow taxis can satisfy all the demand at that particular minute. Furthermore, this linear model was generated by equally distributing the vehicles into the grid and not based on the demand. Therefore, the first two minutes of the model should be ignored. In our future model, we can address this by a different constraint. Another future improvement is the grid size of the model. Our Manhattan grid for the linear programming model is 10×10 to keep the model simple. Each grid is approximately 1,500m × 1,500m versus 300m × 300m which was used in the MDP model. 1,500m is a significantly large size for a grid cell compared to 300m. This creates a significant difference in terms of the demand. We consider no demand if the pick-up and drop-off our at the same grid and we assume that there is no seeking period by the vehicle in the same grid. This would decrease the demand and seeking route significantly. In the future model, we would like to expand to 50×50 with a 60-minute time period, which is $2,500 \times 2,500 \times 60 = 375$ millions data points on one dimension. We must take good care in the set up of the constraints of this model. We can also implement the travel time in the future model that would bring more realistic features to our model. Lastly, having stochastic demand would provide an even more realistic model, especially when traffic accidents occur in real time.

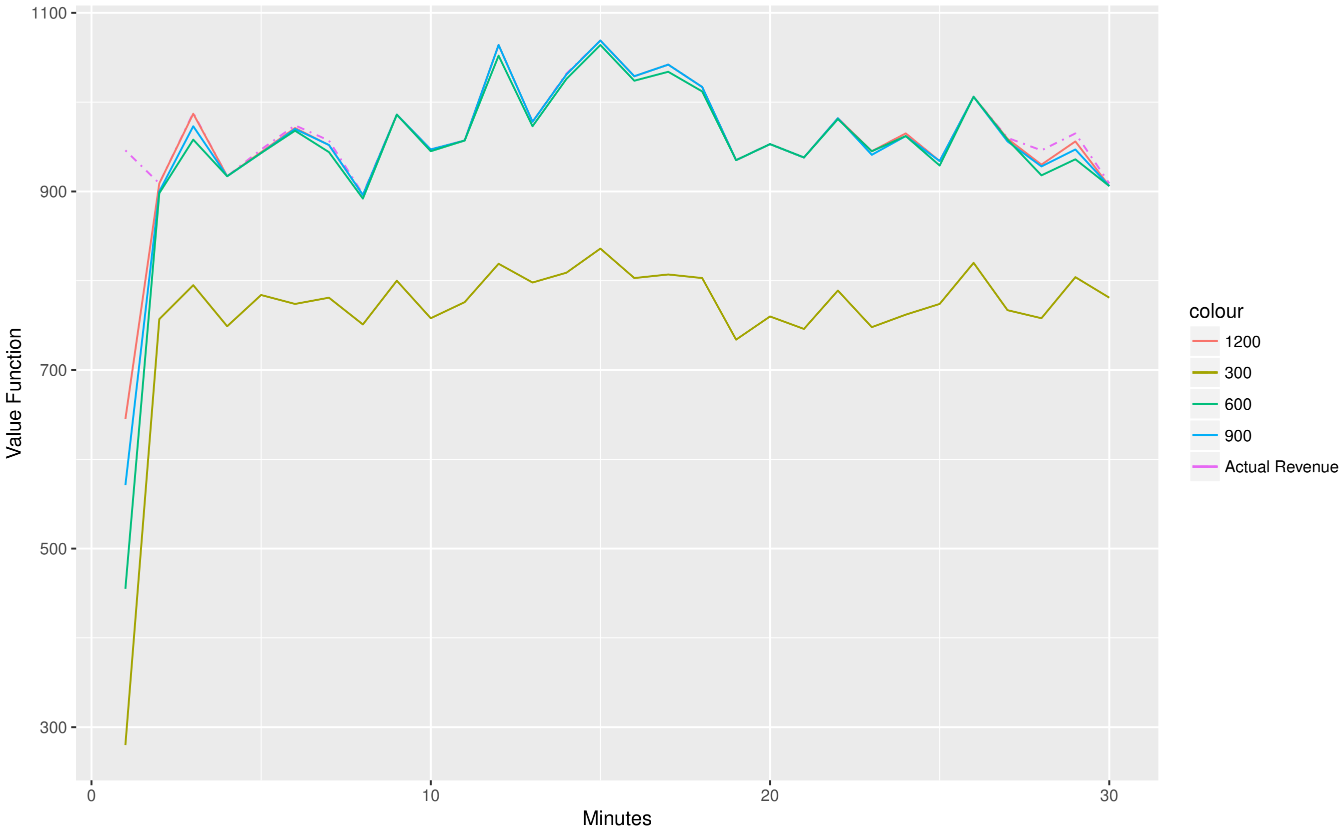

Figure 6. Revenue with the different sizes of the vehicles inventory in 30 minutes.

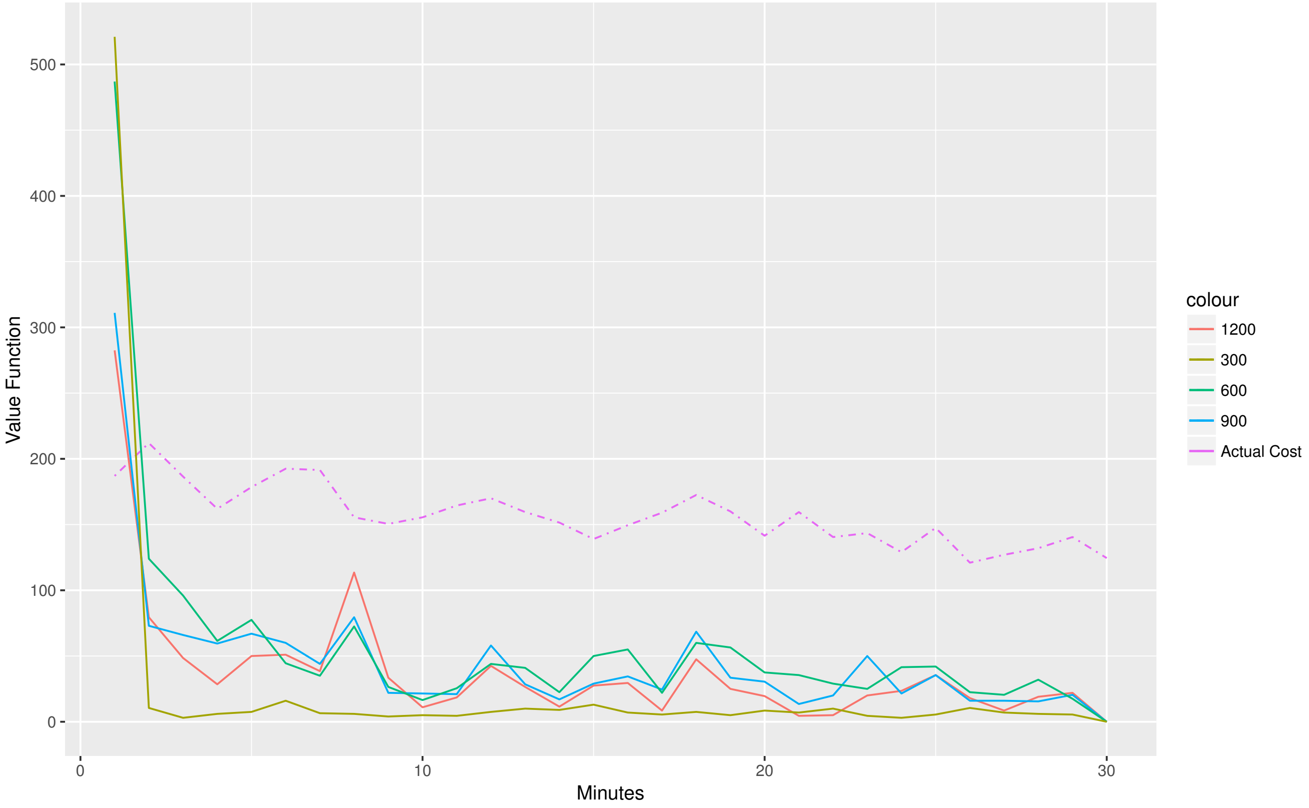

Figure 7. Cost with the different sizes of the vehicles inventory in 30 minutes.

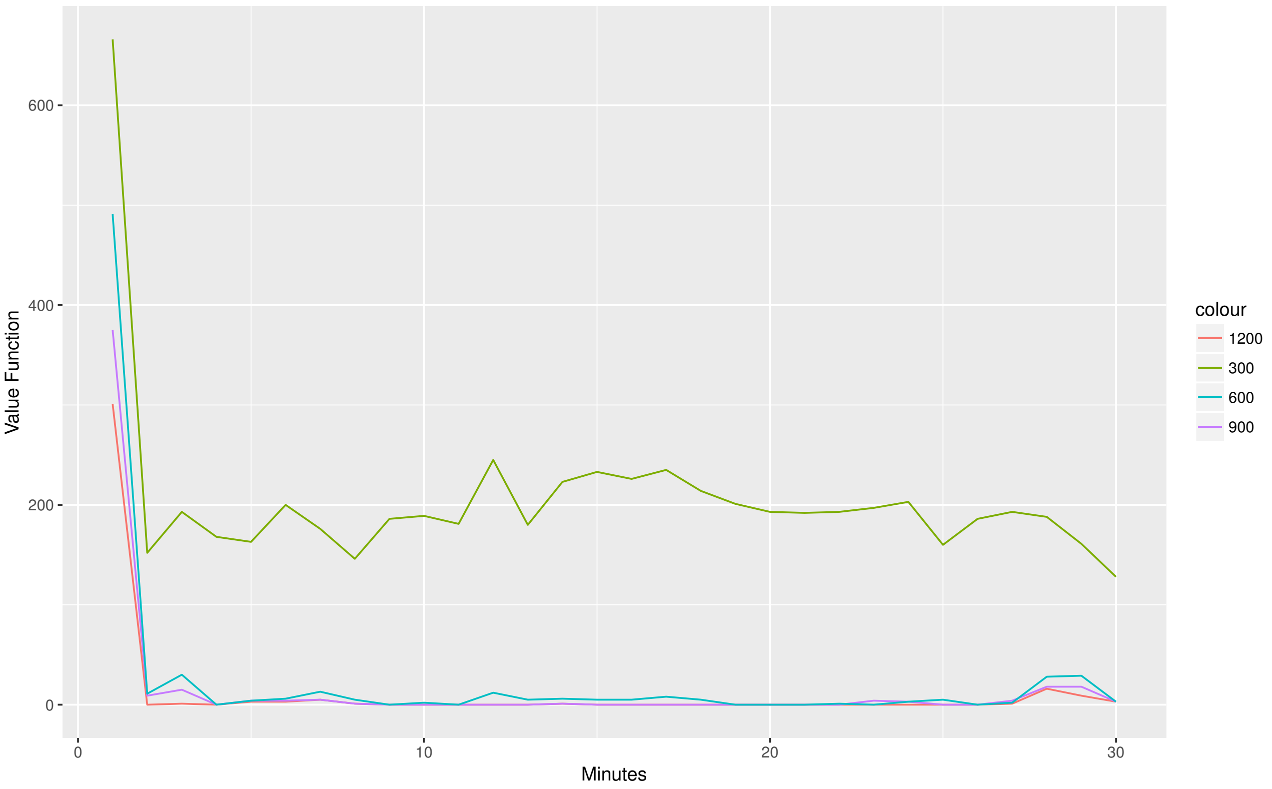

Figure 8. Lost Revenue with with the different sizes of the vehicles inventory in 30 minutes.

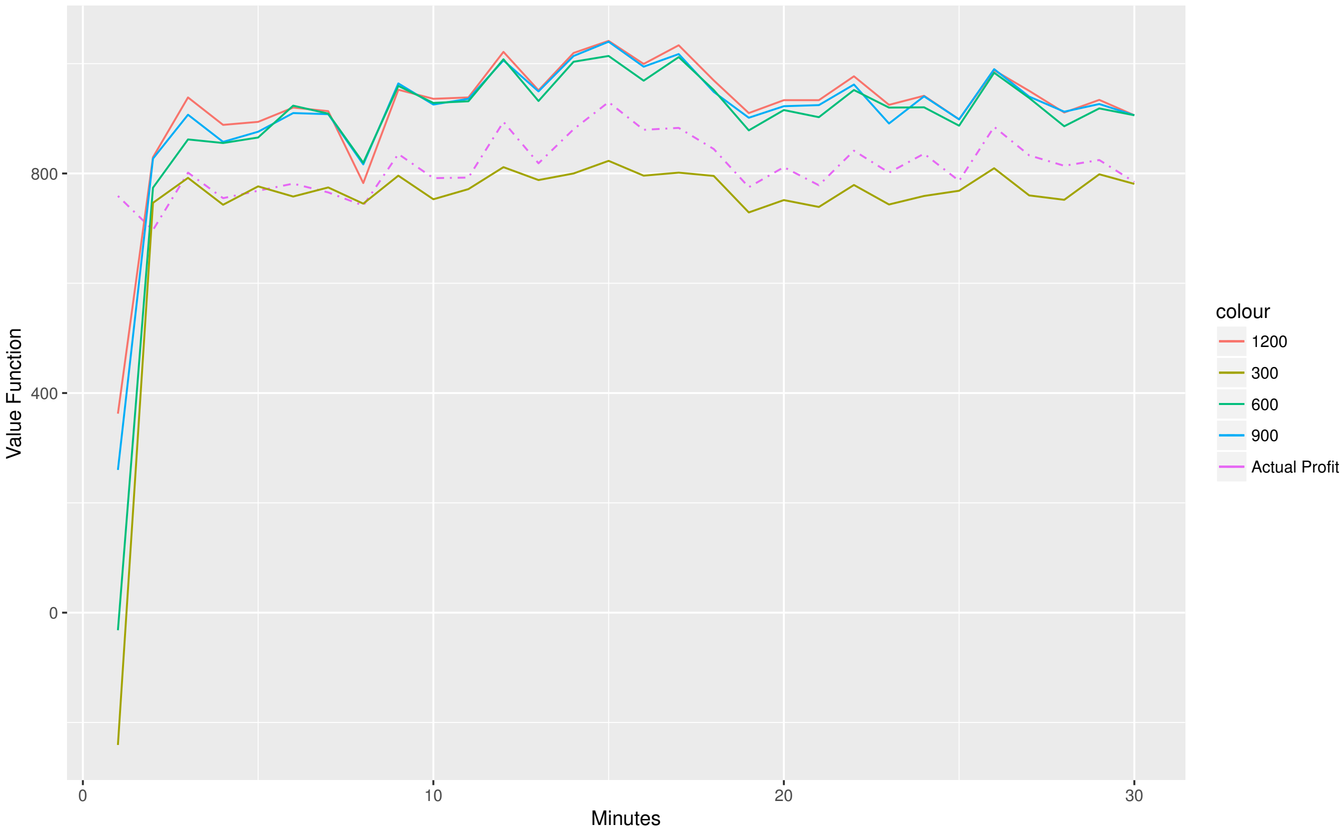

Figure 9. Profit with the different sizes of the vehicles inventory in 30 minutes.

Table IV. Table of Demand, Seeking, Profit and Lost Revenue.

				Percentage Difference of the Profit				Lost Revenue= $C^l \times (D - Optimal^l)$			
Minute	Demand	Seeking	Actual	1200	900	600	300	1200	900	600	300
	$D_{i,j,t}$	$S_{i,j,t}$	Profit	Vehicles	Vehicles	Vehicles	Vehicles	Vehicles	Vehicles	Vehicles	Vehicles
1	488	477	759	-70.71%	-97.94%	-217.61%	-386.10%	301	375	491	666
2	479	485	697	17.36%	17.06%	10.47%	6.86%	0	9	11	152
3	517	478	801.5	15.75%	12.35%	7.27%	-1.19%	1	15	30	193
4	478	439	755	16.25%	12.71%	12.48%	-1.60%	0	0	0	168
5	468	434	768.5	15.10%	13.07%	11.87%	1.04%	3	4	4	163
6	492	452	781.5	16.28%	15.19%	16.66%	-3.05%	3	4	6	200
7	493	471	765.5	17.63%	17.03%	17.14%	1.17%	5	5	13	176
8	473	436	741.5	5.38%	9.63%	9.99%	0.47%	1	1	5	146
9	507	438	835.5	13.09%	14.28%	13.82%	-4.84%	0	0	0	186
10	512	421	791.5	16.73%	15.61%	15.93%	-4.99%	0	0	2	189
11	505	436	792.5	16.87%	16.60%	16.13%	-2.69%	0	0	0	181
12	540	426	894	13.31%	11.79%	11.99%	-9.67%	0	0	12	245
13	494	436	818.5	15.03%	14.82%	12.97%	-3.80%	0	0	5	180
14	536	416	880.5	14.63%	14.09%	13.06%	-9.58%	1	1	6	223
15	529	400	930	11.31%	11.17%	8.64%	-12.21%	0	0	5	233
16	532	418	879.5	12.77%	12.27%	9.68%	-9.97%	0	0	5	226
17	536	410	883	15.71%	14.15%	13.61%	-9.68%	0	0	8	235
18	518	461	844.5	13.78%	11.60%	11.97%	-5.98%	0	0	5	214
19	519	437	775	16.02%	15.09%	12.52%	-6.12%	0	0	0	201
20	506	404	811.5	13.98%	12.80%	12.04%	-7.68%	0	0	0	193
21	514	467	778.5	18.11%	17.15%	14.75%	-5.21%	0	0	0	192
22	518	434	841.5	14.90%	13.36%	12.32%	-7.71%	0	0	1	193
23	506	421	801.5	14.31%	10.58%	13.77%	-7.51%	0	4	0	197
24	519	411	836	11.87%	11.76%	9.62%	-9.66%	0	3	3	203
25	483	410	786.5	13.29%	13.29%	12.01%	-2.32%	0	0	5	160
26	495	386	885	11.00%	11.20%	10.54%	-8.91%	0	0	0	186
27	537	409	833	13.18%	12.07%	11.80%	-9.17%	1	4	2	193
28	519	408	814	11.25%	11.41%	8.47%	-7.92%	16	18	28	188
29	481	407	824.5	12.45%	11.65%	10.79%	-3.20%	9	18	29	161
30	470	371	784.5	14.37%	14.37%	14.37%	-0.45%	3	3	3	128
Average*	505.47	429.97	813.02	14.20%	13.39%	12.30%	-5.02%	1.48	3.07	6.48	189.83

REFERENCES

[1] P. Li, S. Bhulai, and J. van Essen, “Optimization of the revenue of the New York city taxi service using Markov decision processes,” in Proceedings of the 6th International Conference on Data Analytics. IARIA, 2017, pp. 47–52.

[2] N. Taxi, L. Commission et al., “2014 taxicab fact book,” 2014.

[3] Y. Ge, H. Xiong, A. Tuzhilin, K. Xiao, M. Gruteser, and M. Pazzani, “An energy-efficient mobile recommender system,” in Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’10. New York, New York, USA: ACM Press, 2010, p. 899. [Online]. Available: http://dl.acm.org/citation.cfm?doid=1835804.1835918

[4] H. Rong, X. Zhou, C. Yang, Z. Shafiq, and A. Liu, “The rich and the poor: A markov decision process approach to optimizing taxi driver revenue efficiency,” in Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. ACM, 2016, pp. 2329–2334.

[5] J. Yuan, Y. Zheng, L. Zhang, X. Xie, and G. Sun, “Where to find my next passenger,” in Proceedings of the 13th international conference on Ubiquitous computing. ACM, 2011, pp. 109–118.

[6] Y. Zheng, J. Yuan, W. Xie, X. Xie, and G. Sun, “Drive Smartly as a Taxi Driver,” in 2010 7th International Conference on Ubiquitous Intelligence & Computing and 7th International Conference on Autonomic & Trusted Computing. IEEE, oct 2010, pp. 484–486. [Online]. Available: http://ieeexplore.ieee.org/document/5667121/

[7] M. Qu, H. Zhu, J. Liu, G. Liu, and H. Xiong, “A cost-effective recommender system for taxi drivers,” in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’14. New York, New York, USA: ACM Press, 2014, pp. 45–54. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2623330.2623668

[8] D. Zhang, L. Sun, B. Li, C. Chen, G. Pan, S. Li, and Z. Wu, “Understanding Taxi Service Strategies From Taxi GPS Traces,” IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 1, feb 2015, pp. 123–135. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6841047

[9] P. Castro, D. Zhang, and S. Li, “Urban traffic modelling and prediction using large scale taxi GPS traces,” Pervasive Computing, 2012, pp. 57–72.

[10] Y. Ge, C. Liu, H. Xiong, and J. Chen, “A taxi business intelligence system,” in Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’11. New York, New York, USA: ACM Press, 2011, p. 735. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2020408.2020523

[11] B. D. Ziebart, A. L. Maas, A. K. Dey, and J. A. Bagnell, “Navigate like a cabbie: Probabilistic reasoning from observed context-aware behavior,” in Proceedings of the 10th international conference on Ubiquitous computing. ACM, 2008, pp. 322–331.

[12] C.-M. Tseng and C.-K. Chau, “Viability analysis of electric taxis using new york city dataset,” in Proceedings of the Eighth International Conference on Future Energy Systems. ACM, 2017, pp. 328–333.

[13] T. Altshuler, R. Katoshevski, and Y. Shiftan, “Ride sharing and dynamic networks analysis,” arXiv preprint arXiv:1706.00581, 2017.

[14] S. Chawla, Y. Zheng, and J. Hu, “Inferring the Root Cause in Road Traffic Anomalies,” in 2012 IEEE 12th International Conference on Data Mining. IEEE, dec 2012, pp. 141–150. [Online]. Available: http://ieeexplore.ieee.org/document/6413908/

[15] Y. Huang, F. Bastani, R. Jin, and X. S. Wang, “Large scale real-time ridesharing with service guarantee on road networks,” Proceedings of the VLDB Endowment, vol. 7, no. 14, 2014, pp. 2017–2028.

[16] S. Qian, J. Cao, F. L. Mouël, I. Sahel, and M. Li, “Scram: a sharing considered route assignment mechanism for fair taxi route recommendations,” in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015, pp. 955–964.

[17] H. Topaloglu and W. B. Powell, “Dynamic-programming approximations for stochastic time-staged integer multicommodity-flow problems,” INFORMS Journal on Computing, vol. 18, no. 1, 2006, pp. 31–42.

[18] M. S. Maxwell, M. Restrepo, S. G. Henderson, and H. Topaloglu, “Approximate dynamic programming for ambulance redeployment,” INFORMS Journal on Computing, vol. 22, no. 2, 2010, pp. 266–281.

[19] R. Mesa-Arango and S. V. Ukkusuri, “Minimum cost flow problem formulation for the static vehicle allocation problem with stochastic lane

demand in truckload strategic planning," Transportmetrica A: Transport Science, vol. 13, no. 10, 2017, pp. 893–914.

[20] N. Shi, H. Song, and W. B. Powell, "The dynamic fleet management problem with uncertain demand and customer chosen service level," International Journal of Production Economics, vol. 148, 2014, pp. 110–121.

[21] P.-S. You and Y.-C. Hsieh, "A study on the vehicle size and transfer policy for car rental problems," Transportation Research Part E: Logistics and Transportation Review, vol. 64, 2014, pp. 110–121.

[22] H. Li and N. K. Womer, "Solving stochastic resource-constrained project scheduling problems by closed-loop approximate dynamic programming," European Journal of Operational Research, vol. 246, no. 1, 2015, pp. 20–33.

[23] L. Zéphyr and C. L. Anderson, "Integrating storage to power system management," arXiv preprint arXiv:1604.08189, 2016.

[24] D.-P. Song and J. Carter, "Optimal empty vehicle redistribution for hub-and-spoke transportation systems," Naval Research Logistics (NRL), vol. 55, no. 2, 2008, pp. 156–171.

[25] C. Petzold, "How far from true north are the avenues of Manhattan?" 2015. [Online]. Available: http://www.charlespetzold.com/etc/AvenuesOfManhattan/

[26] D. Luxen and C. Vetter, "Real-time routing with openstreetmap data," in Proceedings of the 19th ACM SIGSPATIAL international conference on advances in geographic information systems. ACM, 2011, pp. 513–516.