Dissecting the Learning Curve of Taxi Drivers: A Data-Driven Approach

Menghai Pan, Yanhua Li
Worcester Polytechnics Institute
mpan,yli15@wpi.edu

Xun Zhou
University of Iowa
xun-zhou@uiowa.edu

Zhenming Liu
College of William & Mary
zliu@cs.wm.edu

Rui Song
North Carolina State University
rsong@ncsu.edu

Hui Lu
Guangzhou University
luhui@gzhu.edu.cn

Jun Luo
Lenovo Group Limited
jluo1@lenovo.com

Abstract

Many real world human behaviors can be modeled and characterized as sequential decision making processes, such as taxi driver's choices of working regions and times. Each driver possesses unique preferences on the sequential choices over time and improves their working efficiency. Understanding the dynamics of such preferences helps accelerate the learning process of taxi drivers. Prior works on taxi operation management mostly focus on finding optimal driving strategies or routes, lacking in-depth analysis on what the drivers learned during the process and how they affect the performance of the driver. In this work, we make the first attempt to inversely learn the taxi drivers' preferences from data and characterize the dynamics of such preferences over time. We extract two types of features, i.e., profile features and habit features, to model the decision space of drivers. Then through inverse reinforcement learning we learn the preferences of drivers with respect to these features. The results illustrate that self-improving drivers tend to keep adjusting their preferences to habit features to increase their earning efficiency, while keeping the preferences to profile features invariant. On the other hand, experienced drivers have stable preferences over time.

Index terms— urban computing, inverse reinforcement learning, preference dynamics

1 Introduction

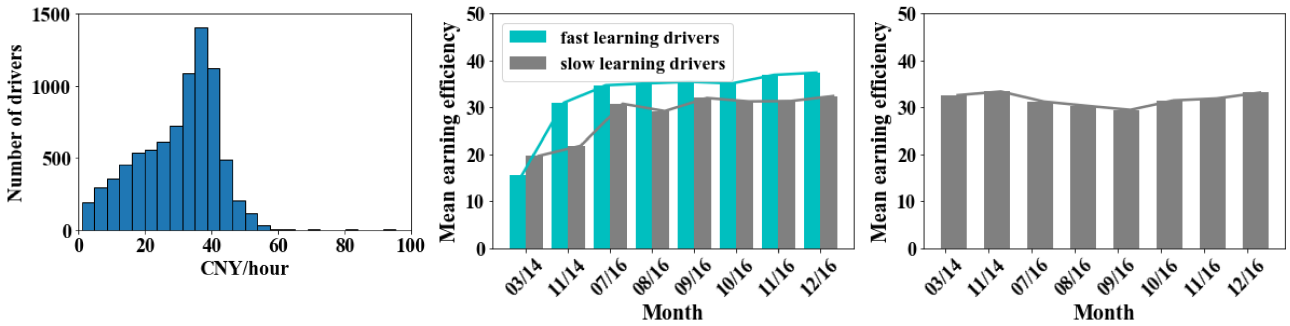
Taxi service is a vital part of the transportation systems in large cities. Improving taxi operation efficiency is a crucial urban management problem, as it helps improve the transportation efficiency of the city and at the same time improves the income of taxi drivers. In the same city, taxi operation efficiency might differ significantly. Fig. 1a shows the earning efficiency (total amount earned normalized by total working time) of different taxi drivers in Shenzhen, China. The top drivers earn 3 to 4 times more money than the bottom drivers.

A major cause of such difference is the difference in working experiences. Fig. 1b shows the growth of earning efficiency of new drivers over years. From March 2014 to December 2016, the new drivers became more experienced and had much higher earning efficiency. During the same time as shown in Fig. 1c, there is no obvious change to the local economy or market, since the average earning efficiency of all the drivers are pretty stable. This shows that drivers are trying to improve their own strategies of looking for passengers based on their increasing knowledge of the city.

However, each driver might learn different knowledge during the learning process, which in turn developed different preferences, when making decisions. For instance, some drivers tend to look for passengers around regions near their homes, and some others might prefer to take passengers from city hubs, e.g., train stations, airport. These preferences might be unique to individual drivers and ultimately lead to differences in earning efficiency. Fig. 1b shows that the "smart" drivers (in blue) improve their earning efficiency faster than "average" drivers and reach a higher level of earning efficiency eventually. Finding what adaptation strategies these "smart" drivers carry could help us understand the learning process of successful drivers and therefore help new drivers to grow faster.

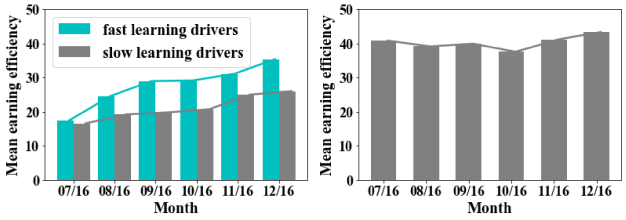
The passenger-seeking behavior of taxi drivers can be modeled as a Markov Decision Process (MDP). Prior work on taxi operation management focused on recommending the optimal policy or routes to maximize the chance of finding passengers or making profit [20, 19, 15, 12]. However, these works only studied how to find the "best" strategies based on data, rather than fundamentally understanding how the drivers learned these strategies over time.

In this work, we make the first attempt to inversely learn the taxi drivers' decision-making preferences, which lead to their choices while looking for passengers. We also study how these preferences evolve



(a) The distribution of earning efficiency in July 2016 (b) The average earning efficiency of new drivers over months (c) The average earning efficiency of all drivers over months

Figure 1: Dynamics of taxi drivers' earning efficiency



(a) New drivers (b) Experienced drivers

Figure 2: Earning efficiency dynamics in 2016

over time and how they help improve the earning efficiency. The results shed lights on "how" the successful drivers became successful, and suggests "smarter" actionable strategies to improve taxi drivers' performances. Our **main contributions** are as follows:

(1) We are the first to employ Inverse Reinforcement Learning to infer the taxi drivers' preferences based on a Markov Decision Process model.

(2) We extract various kinds of interpretable features to represent the potential factors that affect the decisions of taxi drivers.

(3) We infer and analyze the preference dynamics of different groups of taxi drivers.

(4) We conduct experiments with taxi trajectories from more than $17k$ drivers over different time spans. The results verify that each driver has unique preferences to various profile and habit features. The preferences to profile features tend to be stable over time, and the preferences on habit features change over time, which leads to higher earning efficiency.

The rest of the paper is organized as follows. Section 2 motivates and defines the problem. Section 3 details the methodology. Section 4 presents evaluation results. Related works are discussed in Section 5, and the paper is concluded in Section 6.

2 Overview

In this section, we first introduce the motivation of the proposed study, and formally define the problem.

2.1 Motivation

It is a common perception that new drivers gradually learn how to make smart choices as time goes and can improve their working efficiency over time. We verify this perception through data analysis. In Fig. 2a, the average earning efficiency of new drivers who joined in July 2016 increased by up to 25% in 6 months, while in Fig. 2b, the same measure of experienced drivers do not change much. This can be explained by the fact that experienced drivers have learned enough knowledge to make nearly-optimal decisions.

We further noticed that drivers have very different learning curves, which affects ultimately how much earning improvements they can achieve. As previously mentioned, in Fig. 1b, the two colors represent two subgroups of new drivers who joined in March 2014. One group (in blue) are those who became "top" drivers after 2 years with higher earning efficiency, and the other (gray) are the rest of the drivers. Apparently the former had learned more useful knowledge that contributed to their earning improvement. The same diverging trend can be observed among new drivers who joined in July 2016. Fig. 2a shows the comparison of these drivers.

Little is known about what specific knowledge the drivers learned, and which pieces are contributing the most to the earning improvement. Answering these questions would potentially guide and train new drivers to become a quick learner. We consider such "knowledge" as a series of preferences of a driver when making each decision, such as "how frequent to visit the train station", "how far away from home to go when seeking passengers". Specifically, we extract features from the data to represent such decisions a taxi driver might face while working. To achieve the aforementioned goal, in this study we aim to answer two questions: (1) how to recover the preferences of taxi drivers when making these choices, and (2) how these preferences change over time for different groups of drivers.

Problem Definition. In a time interval T_0 i.e.,

(a) Shenzhen road map (b) Map gridding

Figure 3: Shenzhen map data

1 month, given a taxi driver's trajectory data $\tilde{\mathcal{T}}$, and k environmental features $[f_0, f_1, ..., f_k]$, that influence drivers' decision-making process over time, we aim to learn the driver's preference $\theta = [\theta_0, \theta_1, ..., \theta_k]$, i.e., weights to features when the driver makes decisions. Secondly, for a long time horizon, with multiple time intervals $[T_0, T_1, ..., T_m]$, we analyze the evolution pattern of the driver's preferences over time.

2.2 Data Description Our analytical framework takes two urban data sources as input, including (1) taxi trajectory data and (2) road map data. For consistency, both datasets are collected in Shenzhen, China in 2014 and 2016.

The **taxi trajectory data** contains GPS records collected from taxis in Shenzhen, China during 2014 and 2016. There were in total 17, 877 taxis equipped with GPS sets, where each GPS set generates a GPS point every 40 seconds on average. Overall, a total of 51,485,760 GPS records are collected on each day, and each record contains five key data fields, including taxi ID, time stamp, passenger indicator, latitude and longitude. The passenger indicator field is a binary value, indicating if a passenger is aboard or not.

The **Road map data** of Shenzhen covers the area defined between 22.44° to 22.87° in latitude and 113.75° to 114.63° in longitude. The data is from OpenStreetMap [1] and has 21,000 roads of six levels.

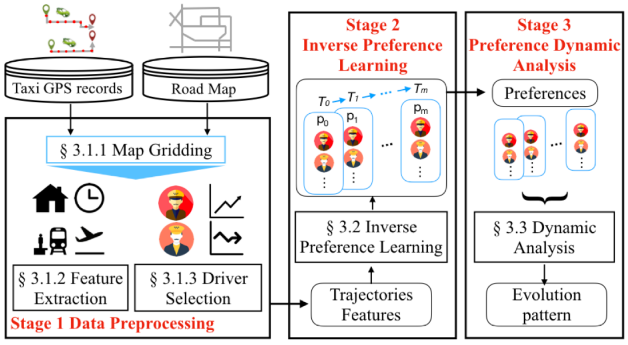


Figure 4: Solution Framework

3 Methodology

Fig. 4 outlines our solution framework, which takes two sources of urban data as inputs and contains three key analytical stages: (1) data preprocessing, (2) inverse preference learning and (3) preference dynamic analysis.

3.1 Data Preprocessing

3.1.1 Map and Time Quantization We use a standard quantization trick to reduce the size of the location space. Specifically, we divide the study area into equally-sized grid cells with a given side-length s in latitude and longitude. Our method has two advantages: (i) we have the flexibility to adjust the side-length to achieve different granularities, and (ii) it is easy to implement and highly scalable in practice [9, 8]. Fig. 3b shows the actual grid in Shenzhen, China with a side-length $l = 0.01°$ in latitude and longitude. Eliminating cells in the ocean, those unreachable from the city, and other irrelevant cells gives a total of 1158 valid cells. We divide each day into five-minute intervals for a total of 288 intervals per day. A spatio-temporal region r is a pair of a grid cell s and a time interval t . The trajectories of drivers then can be mapped to sequences of spatio-temporal regions.

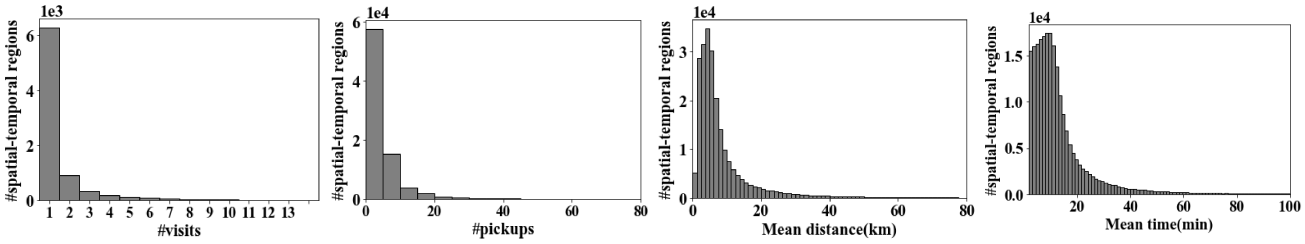
3.1.2 Feature Extraction Taxi drivers make hundreds of decisions throughout their work shifts (e.g., where to find the next passenger, and when to start and finish working in a day). When making a decision, they instinctively evaluate multiple factors (i.e., features) related to their current states and the environment (i.e., the current *spatio-temporal region*). For example, after dropping off a passenger, a driver may choose to go back to an area that she is more familiar with, or a nearby transport station, e.g., airport, train station. Here, we extract key features the drivers use to make their decisions.

Note in our framework, each feature is defined as a numeric characteristic of a specific spatio-temporal region, which may or may not change from driver to driver. For example, let f_r represent the average number of taxi pickups in history in location s during time slot t. Apparently the value of feature f_r is the same for every driver. However, another feature g_r at r could be the distance from s to the home of the driver. The value of this feature varies from driver to driver, depending on their home locations. However, it does not change over time.

The features we extract can be roughly categorized by *profile features* and *habit features*, as detailed below.

Profile Features. Each driver has unique personal (or profile) characteristics, such as home location, daily working schedule (time duration), and preferred geographic area. For each spatio-temporal region, we build the profile features. Here, we extract 4 profile features:

P1: Visitation Frequency. This group of features repre-



(a) *P1:* Number of visits (b) *H1:* Number of pickups (c) *H2:* Mean trip distance (d) *H3:* Mean trip time

Figure 5: Statistical distributions of features

sents the numbers of daily visits to different regions of a driver as extracted from the historical data. Fig. 5a shows the distribution of visitation frequency to different regions of an arbitrarily chosen driver. Here, visitation frequencies vary significantly across regions.

P2: Distance to Home. Each taxi driver has a home location, which can be extracted from their GPS records. This feature characterizes the distance (in miles on the road network) from the current location to the driver's home location. Different drivers may have different preferences in working close to their homes or not.

P3 & P4: Time from Start & Time to Finish. Taxi drivers typically work according to consistent starting and finishing times. We construct two features to characterize the differences of the current time from the regular starting and finishing time.

Habit Features. These represent the habits of the drivers, which are typically governed by experience (e.g., remaining near the train station instead of traveling around to find passengers). We extract 6 habit features.

H1: Number of pickups. This feature characterizes the demands in a cell during a time interval, and is extracted and estimated using the historical trajectories from all drivers. The distribution on the numbers of pickups is shown in Fig. 5b.

H2 & H3: Average Trip Distance & Time. These features represent average distance and travel time of passenger trips starting from a particular spatio-temporal region. A driver's preference to these features characterize how much the driver prefers long vs short distance passenger trips. The distribution of these features across spatio-temporal features are shown in Fig. 5c and Fig. 5d, respectively.

H4: Traffic Condition. This feature captures the average traffic condition based on the time spent by a driver in each spatio-temporal region. A long travel time implies traffic congestion. The preference of drivers over this feature represents how much drivers would like avoid the traffic.

H5 & H6: Distance to Train Station & Airport. These features reflect the distances from the current cell to Shenzhen train station and airport, respectively.

3.1.3 Driver Selection. Different drivers have different earning efficiencies as shown in Fig. 1a. Below, we describe the criteria we use to select drivers.

We estimate the earning efficiency of each driver in different time periods from their historical data. The earning efficiency r_e is defined the average per hour income (i.e., in eq.3.1).

$$r_e = \frac{E}{t_w}, \tag{3.1}$$

where E is the income in the whole sampling time span, span (e.g., per month), and t_w represents the driver's working time.

Driver selection criterion: We select drivers with the highest earnings, because the preference learning algorithms require the input data to be generated by the converged policy (see more details in Sec 3.2). We note that drivers with high earning efficiencies are likely the most experienced (i.e., they use converged policies to make decisions).

3.2 Inverse Preference Learning

This section explains our inverse learning algorithm for extracting drivers' decision-making process. We use a Markov Decision Process (MDP) to model drivers' sequential decision-making and relative entropy inverse reinforcement learning (REIRL) to learn their decision-making preferences.

3.2.1 Markov Decision Process. A Markov Decision Process(MDP) [3] is defined by a 5-tuple $\langle S, A, T, \gamma, \mu_0, R\rangle$ so that

- S is a finite set of states and A is a finite set of actions,
- T is the probabilistic transition function with $T(s'|s,a)$ as the probability of arriving at state s' by executing action a at state s,
- $\gamma \in (0,1]$ is the discount factor[1],
- $\mu_0 : S \to [0,1]$ is the initial distribution, and
- $R : S \times A \to$ is the reward function.

[1] Without loss of generality, we assume $\gamma = 1$ in this work, and it is straightforward to generalize our results to $\gamma \neq 1$.

A randomized, memoryless policy is a function that specifies a probability distribution on the action to be executed in each state, defined as $\pi : S \times A \to [0, 1]$. We use $\tau = [(s_0, a_0), (s_1, a_1), \ldots, (s_L, a_L)]$ to denote a trajectory generated by MDP. Here L is the length of trajectory. We model the decision-making process of taxi drivers with MDP as follow:

- State: a spatio-temporal region, specified by a geographical cell and a time slot.
- Action: traveling from the current cell to one of the eight neighboring cells, or staying in the same cell.
- Reward: the inner product of the preference function (as a vector) θ and the feature vector $\mathbf{f}$ on each state-action pair.

Fig. 6 shows an example of trajectory in the MDP: a driver starts in state s_0 with the taxi idle, and takes the action a_0 to travel to the neighboring cell S_1. After two steps, the driver reaches state S_2, where she meets a passenger. The destination of the new trip is cell S_3. The trip with the passenger is a transition in the MDP from S_2 to S_3.

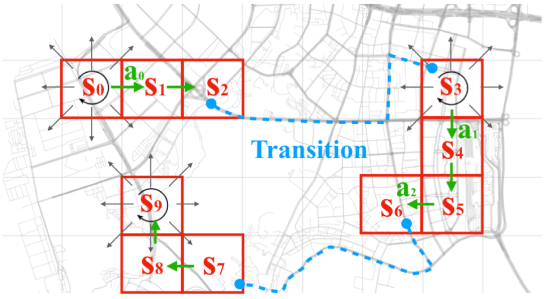


Figure 6: MDP of taxi driver's decision making process

3.2.2 Inverse Preference Learning. Given the observed trajectory set $\tilde{\mathcal{T}}$ of a driver and the features extracted on each state-action pair (s, a), the inverse preference learning stage aims to recover a reward function (i.e., preference vector θ) under which the observed trajectories have the highest likelihood to be generated [13]. Various inverse reinforcement learning approaches, e.g., Apprenticeship learning [2], Maximum Entropy IRL [24], Bayesian IRL [14] and Relative Entropy IRL [4], have been proposed in the literature. Our problem possesses two salient characteristics: (i) the state space is large. We have 1158 cells and 288 time intervals. Therefore, the total number of states is $1158 \times 288 \approx 330$k, and (ii) the transition probability is hard to measure because in part of the large state space issue. Therefore, we adopt a model-free IRL approach, namely, relative entropy IRL [4] that does not require estimating transition probabilities and is more scalable than other alternatives.

The optimization problem. Let $\mathcal{T}$ denote the set of all possible trajectories of the driver decision-making MDP, outlined in Sec 3.2.1. For any $\tau \in \mathcal{T}$, denote $P(\tau)$ as the trajectory distribution induced by the taxi driver's ground-truth policy, and $Q(\tau)$ as the trajectory distribution induced by a base policy. The Relative Entropy between $P(\tau)$ and $Q(\tau)$ (in eq.3.2) characterizes as the distribution difference between $P(\tau)$ and $Q(\tau)$.

$$H(P\|Q) = \sum_{\tau \in \mathcal{T}} P(\tau) \ln \frac{P(\tau)}{Q(\tau)}. \tag{3.2}$$

The driver's trajectory distribution is governed by the driver's preference θ, thus is a function of θ, i.e., $P(\tau|\theta)$. The relative entropy IRL aims to find a reward function θ, that minimizes the relative entropy in eq.3.2 and matches the trajectory distribution to the observed trajectory data.

P1: Relative Entropy IRL Problem:

$$\min_{\theta}: \quad H(P(\theta)\|Q) = \sum_{\tau \in \mathcal{T}} P(\tau|\theta) \ln \frac{P(\tau|\theta)}{Q(\tau)}, \tag{3.3}$$

$$\text{s.t.:} \quad |\sum_{\tau \in \mathcal{T}} P(\tau|\theta) f_i^\tau - \hat{f}_i| \le \epsilon_i, \forall i \in \{1, ..., k\}, \tag{3.4}$$

$$\sum_{\tau \in \mathcal{T}} P(\tau|\theta) = 1, \tag{3.5}$$

$$P(\tau|\theta) \ge 0, \qquad \forall \tau \in \mathcal{T}, \tag{3.6}$$

where i is the feature index, and f_i^τ is the i's feature count in trajectory τ, and $\hat{f}_i = \sum_{\tau \in \tilde{\mathcal{T}}} f_i^\tau / |\tilde{\mathcal{T}}|$ is the feature expectation over all observed trajectories in $\tilde{\mathcal{T}}$. ϵ_i is a confidence interval parameter, which can be determined by the sample complexity (the number of trajectories) via applying a Hoeffding's bound. The constraint eq.3.4 ensures that the recovered policy matches the observed data. The constraints eq.3.5–eq.3.6 guarantees the $P(\tau|\theta)$'s are non-negative probabilities, thus sum up to one.

Solving P1. The function $Q(\tau)$ and $P(\tau|\theta)$ can be decomposed as

$$Q(\tau) = T(\tau)U(\tau) \text{ and } P(\tau|\theta) = T(\tau)V(\tau|\theta),$$

where $T(\tau) = \mu_0(s_0) \prod_{t=1}^{K} T(s_t|s_{t-1}, a_{t-1})$ is the joint probability of the state transitions in τ, for $\tau = [(s_0, a_0), (s_1, a_1), \cdots, (s_K, a_K)]$, with $\mu_0(s_0)$ as the initial state distribution. $U(\tau)$ (resp. $V(\tau|\theta)$) is the joint probability of the actions conditioned on the states in τ under driver's policy π_θ (resp. a base policy π_q). As a result, eq.3.3 can be written as follows.

$$H(P(\theta)\|Q) = \sum_{\tau \in \mathcal{T}} P(\tau|\theta) \ln \frac{V(\tau|\theta)}{U(\tau)}. \tag{3.7}$$

Moreover, when $\pi_q(a|s)$ at each state s is uniform distribution, e.g., $\pi_q(a|s) = 1/|A_s|$, with A_s as the set of actions at state s, the problem **P1** is equivalent to maximizing the causal entropy of $P(\tau|\theta)$, i.e., $\sum_{\tau \in \mathcal{T}} P(\tau|\theta) \ln V(\tau|\theta)$, while matching $P(\tau|\theta)$ to the observed data [23]. Following the similar process outlined in [4], **P1** can be solved by a gradient descent approach, with the step-wise updating gradient as follows.

$$\nabla g(\theta) = \hat{f}_i - \frac{\sum_{\tau \in \mathcal{T}^\pi} \frac{U(\tau)}{\pi(\tau)} \exp(\sum_{j=1}^k \theta_i f_j^\tau)}{\sum_{\tau \in \mathcal{T}^\pi} \frac{U(\tau)}{\pi(\tau)} \exp(\sum_{j=1}^k \theta_i)} - \alpha_i \epsilon_i, \tag{3.8}$$

where $\alpha_i = 1$ if $\theta_i \leq 0$ and $\alpha_i = -1$ otherwise. $\mathcal{T}^\pi$ is a set of trajectories sampled from $\tilde{\mathcal{T}}$ by an executing a given policy π. $U(\tau)$ is the joint probability of taking actions conditioned on the states in a observed trajectory τ, induced by uniform policy $\pi_q(a|s) = 1/|A_s|$. See Algorithm 1 for our IRL algorithm.

Algorithm 1 Relative Entropy IRL

Input: Demonstrated trajectories $\tilde{\mathcal{T}}$, feature matrix F, threshold vector ϵ, learning rate α, and executing policy π.
Output: Preference vector θ.
1: Randomly initialize preference vector θ.
2: Sample a set of trajectories. $\mathcal{T}^\pi$ using π.
3: Calculate feature expectation vector $\hat{f}$.
4: **repeat**
5: Calculate each feature count f_i^τ.
6: Calculate gradient $\nabla g(\theta)$ using Eq 3.8.
7: Update $\theta \leftarrow \theta + \alpha \nabla g(\theta)$.
8: **until** $\nabla g(\theta) < \epsilon$.

3.3 Preference Dynamic Analysis. Using Algorithm 1, we can inversely learn the preference θ for each driver, during each time interval (e.g., a month) over time, and obtain a sequence of preference vectors $\{\theta_1, \cdots, \theta_N\}$. For each driver, we can conduct hypothesis testing to examine if the change of the preference vectors over months is significant or not. We denote the preference vector learned for taxi driver p in period T_i as θ_i^p, and that in period T_j as θ_j^p. Then, we can obtain two preference vector sample sets in i-th and j-th months as S_i and S_j over a group of n drivers as follow:

$$S_i = \{\theta_i^1, \theta_i^2, ..., \theta_i^n\}, \tag{3.9}$$

$$S_j = \{\theta_j^1, \theta_j^2, ..., \theta_j^n\}. \tag{3.10}$$

With S_i and S_j, we will examine if the entries in preference vectors changed significantly or not from the i-th to j-th month, using two-sample t-test [17]. For each feature f_m, the null hypothesis is that the difference between the m-th entry of each θ_i^p in S_i and θ_j^p in S_j equals 0, which means drivers' preference to feature f_m does not change significantly from the i-th month to the j-th month. Otherwise, the alternative hypothesis indicates a significant change. Taking the difference between S_i and S_j as $\Delta S_{ij} = \{\Delta\theta_{ij}^1, \Delta\theta_{ij}^2, ..., \Delta\theta_{ij}^n\} = \{\theta_i^1 - \theta_j^1, \theta_i^2 - \theta_j^2, ..., \theta_i^n - \theta_j^n\}$. The t-test statistics of the m-th entry is as follow.

$$t_{ij}(m) = \frac{Z}{s} = \frac{\Delta\overline{\theta}_{ij}(m) - \mu}{\delta/\sqrt{n}}. \tag{3.11}$$

where μ is the sample mean, n is the sample size and δ if the sample square error. The t-distribution for the test can be determined given the degree of freedom $n-1$. Given a significance value $0 < \alpha < 1$, we can get a threshold of the t value t_α in the t-distribution. Then if $t_{ij}(k) > t_\alpha$, the null hypothesis should be rejected with significance α, otherwise, we can accept the null hypothesis with significance α. Usually, we set $\alpha = 0.05$, which also means the confidence of the test is $1 - \alpha = 0.95$.

4 Experiments

In this section, we conduct experiments with real world taxi trajectory data to learn the preferences of different groups of taxi drivers, and analyze the preference evolution patterns for each group.

4.1 Experiment settings When analyzing the temporal dynamics of the drivers' decision-making preferences, the null hypothesis is that the difference between the preferences in two time periods is not significant. The alternative hypothesis is the temporal preference difference is significant. We choose the t-test significance value $\alpha = 0.05$.

Driver Group Selection. We aim to analyze how taxi drivers' decision making preferences evolve over time. For each month, we select 3000 drivers with the highest earning efficiency. The intuition is that these drivers are likely more experienced drivers, thus with near-optimal policies, under maximum causal entropy principle [24]. To evaluate the preference change across two months, i.e., the i-th and j-th months, we find those drivers from those experience drivers, who also show up in both months for our study. For example, in 07/2016 and 12/2016, there are 2151 experienced drivers in common. Then, we calculate the difference of earning efficiency of each driver in the two months. Fig. 7 shows the gap distribution in 07/2016 and 12/2016. We will choose two groups of drivers for preference dynamics analytics based on the drivers' earning efficiency gaps.

- Group #1 (Self-improving Drivers): 200 drivers whose earning efficiencies increase the most.

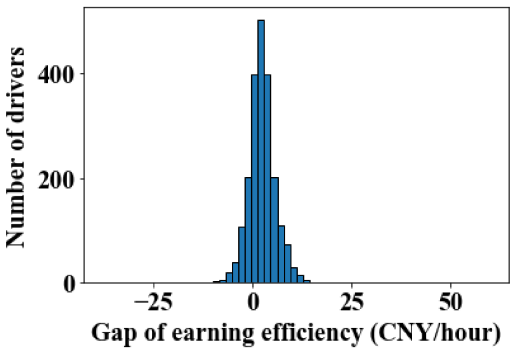


Figure 7: Earning efficiency gap distribution

- Group #2 (Stabilized Drivers): 200 drivers whose earning efficiency gaps are small, i.e., close to 0.

Experiment Plan. We use 12 months trajectory data across three years of time span for our study, i.e., 07/2014–12/2014, and 07/2016–12/2016. We evaluate the preference dynamics across months pairs. First, we setup the month 07/2016 as the base month, and compare the preferences of drivers in Group #1 and Group #2, with that of 5 subsequent months (08/16, 09/16, 10/16, 11/16, and 12/16), respectively. Then, we compare the preferences of drivers in Group #1 and Group #2 in the same month in 2014 vs 2016.

4.2 Preference dynamics analysis. Now we present the results on analyze the preference dynamics of two driver groups over time.

Results for Group #1. The table on the top in Fig. 8 shows the t-values obtained for comparing preferences (with respect to each feature) in 07/16 to that of 08/16, 09/16, 10/16, 11/16, and 12/16, respectively. For these self-improving drivers, The boxes of failed tests are marked with red color, and the corresponding t values. First, with a time span of less than three months, the preferences do not show any significant change. However, when the time space is larger than three months, preferences to some habit features changed significantly, including *H1: Number of pickups*, *H3: average trip time* and *H4: traffic condition*. This makes sense, since over time, the self-improving drivers tend to learn the knowledge of where the demands, low traffic, long trip orders are. On other hand, the preferences to all four profile features and other habit features stay unchanged over the half a year.

Results for Group #2: The lower table in Fig. 8 shows the t-values obtained for preference comparison of drivers in Group #2. Clearly, the preferences to all profile and habit features stay unchanged over the half a year, which means that these stabilized drivers have kept the same strategy of finding passengers in the half a year. This is consistent with their unchanged earning efficiencies over time.

Group #1 (Self-improving Drivers)

	P1	P2	P3	P4	H1	H2	H3	H4	H5	H6
Aug	-1.09	0.82	-0.27	0.61	0.15	-0.44	-0.03	0.93	1.23	-1.33
Sep	0.70	0.79	0.26	-0.36	-1.88	1.11	0.66	0.70	-0.33	-0.43
Oct	-0.18	0.08	-0.48	0.51	0.02	-1.66	0.89	-1.30	0.19	1.12
Nov	1.75	-0.96	-0.10	-1.63	-2.20	0.58	1.39	2.80	1.30	-0.34
Dec	-0.43	0.43	-0.32	-0.10	-2.51	-0.28	2.22	2.11	0.01	-0.34

Group #2 (Stabilized Drivers)

	P1	P2	P3	P4	H1	H2	H3	H4	H5	H6
Aug	-1.05	-1.23	1.30	-1.28	0.71	-0.66	-0.28	-1.94	-0.47	1.21
Sep	-0.08	-0.23	1.44	-0.85	0.09	0.83	-0.99	-0.98	-0.63	0.24
Oct	0.04	-1.74	1.87	-1.20	0.84	0.25	-0.63	-1.48	-0.19	-0.40
Nov	-0.62	0.77	-0.11	-0.95	1.05	-0.25	-1.50	-1.06	0.44	0.05
Dec	0.88	-1.13	1.89	0.36	-0.21	-0.36	0.57	-0.76	-1.27	-0.11

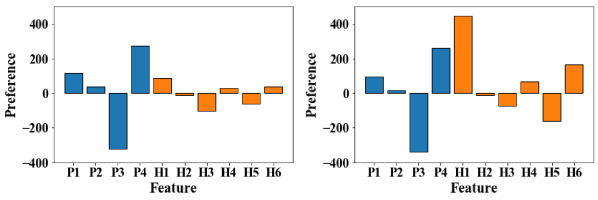
Figure 8: Preference dynamics within a year

Similar results are also obtained when evaluating the preference changes across years, i.e., between the same month in 2014 vs 2016. We omitted the results here due to the limited space.

4.3 Case Study. To further understand the preference dynamics, we look into individual drivers show case how the preference and working behaviors evolve over time. Here we show one randomly selected driver from Group #1. Let us call him "John". John's earning efficiency grew from 41.84 CNY/hour to 52.24 CNY/hour from 07/16 to 12/16. His preferences in both months are listed in Fig. 9a -9b. Clearly, the preferences to the profile features remain unchanged, while the preferences to some habit features, such as *H1 Number of pickups*, *H5&H6 Distance to Train Station & Airport* changed. When we look into John's driving behaviors, it matches the preference change perfectly.

Preference change to H1. The preference change to feature H1 indicates that John increased his preference to areas with high volume of pickup demands. Fig. 10(a)-(b) shows the distribution of trajectories when the taxi was idle in the morning rush hours in 07/16 and 12/18, respectively. Fig. 10(c)-(d) shows the all taxi pickup demand distributions in the morning rush hours in 07/16 and 12/16, respectively. The city-wide demand distribution does not change. However, during the morning rush hours, John changed his strategy from 07/16 to 12/16, i.e., to look for passengers from the high demand areas. This is consistent to the preference change to feature H1 (number of pickups).

Preference change to H5. The preference change to feature H5, i.e., distance to train station, is also significant. The negative preference indicates John prefers to be closer to the train station to look for passengers. Over time this preference became stronger.



(a) Preference in 07/16 (b) Preference in 12/16

Figure 9: The decision-making preferences of John

To explain this phenomenon, we highlighted the train station in Fig. 10(a)-(b). The statistics we obtained from John's trajectory data showed that the percentage of order received near the train station increased from 11.93% in 07/16 to 14.21% in 12/16, which is consistent with the preference change.

4.4 Takeaways and Discussions. From our studies on a large amount of taxi trajectory data spanning for 3 years, *we made the first ever report on how real world taxi drivers make decisions when looking for passengers, and how their preferences evolve over time.* Overall, two key takeaways are summarized as follows.

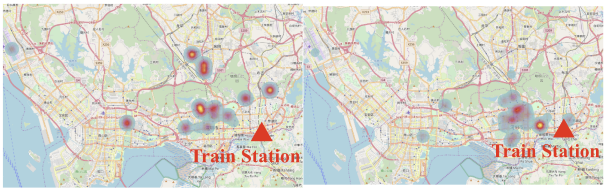
1. Each driver has its unique preferences to their profile features, which tend to be stable over time.
2. Drivers while learning the environments, may change their preferences to habit features.

Our findings can be potentially utilized to assist and guide taxi drivers to improving their earning efficiencies. For example, for those slow learning drivers, by learning their preferences, especially, the preferences to habit features, we can diagnose which knowledge in terms of the features they are lacking, e.g., not familiar with the high demand regions. As a result, some guiding messages, may be sent directly to the drivers about such information, to assist the drivers to improve a better policy faster.

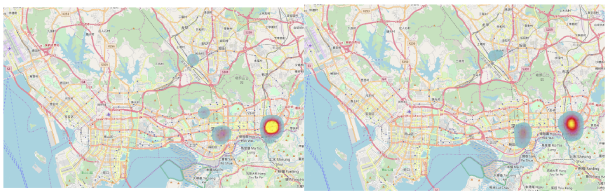
5 Related Works

Taxi operating strategies (e.g., dispatching, passenger seeking), and driver behavior analysis have been extensively studied in recent years due to the emergence of the ride-sharing business model and urban intelligence. However, to the best of our knowledge, *we make the first attempt to employ inverse reinforcement learning to analyze the preference dynamics of taxi drivers.* Related works to our study are summarized below.

Urban Computing is a general research area which integrates urban sensing, data management and data analytic together as a unified process to explore, analyze and solve crucial problems related to people's everyday life [8, 11, 5, 10, 22, 7, 12, 18]. In particular, a group of work study taxi operation management, such as dispatching [16, 6] and passenger seeking [21, 19, 20],



(a) Heatmap of the trajectory in July, 2016 (b) Heatmap of the trajectory in December, 2016



(c) Heatmap of pickups in July, 2016 (d) Heatmap of pickups in December, 2016

Figure 10: A case study

aiming at finding an optimal actionable solution to improve the performance/revenue of individual taxi drivers or the entire fleet. [15] solved the passenger seeking problem by giving direction recommendations to drivers. However, all of these works focus on finding "what" are the best driving strategies (as an optimization problem), rather than finding "why" and "how" good drivers make these decisions. By contrast, our work focuses on analyzing the evolving preferences of good drivers, that drive them to make better and more profitable decisions.

Inverse Reinforcement Learning(IRL) aims to recover the reward function under which the expert's policy is optimal from the observed trajectories of an expert. There are various of IRL methods, for example, [13] found that there are a class of reward functions that can lead to the same optimal policy, and it proposed a strategy to select a reward function. However, this method is not proper to analyze human behaviors because it uses the deterministic policy in the Markov Decision Process while human decisions tend to be non-deterministic. And [24] proposed a IRL method by maximizing the entropy of the distribution on state-actions under the learned policy. Although this method can employ stochastic policy, the computation efficiency is not friendly to large scale state space, and it requires the information of the model. In this paper, we employ Relative Entropy IRL [4] which is model-free and employs softmax policy. Our work, compared to the above related work, is the first to apply IRL to study the evovling driving preferences of taxi drivers.

6 Conclusion

In this paper, we made the first attempt to employ inverse reinforcement learning to analyze the preferences of taxi drivers when making sequences of decisions to

look for passengers. We further studied how the drivers' preferences evolve over time, during the learning process This problem is critical to helping new drivers improve performance fast. We extracted different types of interpretable features to represent the potential factors that affect the decisions of taxi drivers and inversely learned the preferences of different groups of drivers. We conducted experiments using large scale taxi trajectory datasets, and the results demonstrated that drivers tend to improve their preferences to habits features to gain more knowledge in the learning phase and keep the preferences to profile features stable over time.

7 Acknowledgements

Yanhua Li and Menghai Pan were supported in part by NSF grants CNS-1657350 and CMMI-1831140, and a research grant from DiDi Chuxing Inc. Xun Zhou was partially supported by NSF grant IIS-1566386. Zhenming Liu was supported by NSF grant IIS-1755769. Rui Song's research was partially supported by NSF grant DMS-1555244.

References

[1] OpenStreetMap. `http://www.openstreetmap.org/`.

[2] P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1. ACM, 2004.

[3] R. Bellman. A markovian decision process. *Indiana Univ. Math. J.*, 6:679–684, 1957.

[4] A. Boularias, J. Kober, and J. Peters. Relative entropy inverse reinforcement learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 182–189, 2011.

[5] Y. Ding, Y. Li, K. Deng, H. Tan, M. Yuan, and L. M. Ni. Detecting and analyzing urban regions with high impact of weather change on transport. *IEEE Transactions on Big Data*, 2016.

[6] J. Yuan, Y. Zheng, L. Zhang, X. Xie. T-Finder: A Recommender System for Finding Passengers and Vacant Taxis. *IEEE Transactions on Knowledge and Data Engineering*, 25(10):2390–2403, 2013.

[7] A. V. Khezerlou, X. Zhou, L. Li, Z. Shafiq, A. X. Liu, and F. Zhang. A traffic flow approach to early detection of gathering events: Comprehensive results. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(6):74, 2017.

[8] Y. Li, J. Luo, C.-Y. Chow, K.-L. Chan, Y. Ding, and F. Zhang. Growing the charging station network for electric vehicles with trajectory data analytics. In *ICDE*, 2015.

[9] Y. Li, M. Steiner, J. Bao, L. Wang, and T. Zhu. Region sampling and estimation of geosocial data with dynamic range calibration. In *ICDE*, 2014.

[10] C. Liu, K. Deng, C. Li, J. Li, Y. Li, and J. Luo. The optimal distribution of electric-vehicle chargers across a city. In *ICDM*. IEEE, 2016.

[11] B. Lyu, S. Li, Y. Li, J. Fu, A. C. Trapp, H. Xie, and Y. Liao. Scalable user assignment in power grids: a data driven approach. In *SIGSPATIAL GIS*. ACM, 2016.

[12] M. Qu, H. Zhu, J. Liu, G. Liu, H. Xiong. A Cost-Effective Recommender System for Taxi Drivers. In *The 20th International Conference on Knowledge Discovery and Data Mining (SIGKDD'14)*, pages 45–54, New York, NY, 2014. ACM.

[13] A. Y. Ng, S. J. Russell, et al. Algorithms for inverse reinforcement learning.

[14] D. Ramachandran and E. Amir. Bayesian inverse reinforcement learning. *Urbana*, 51(61801):1–4, 2007.

[15] 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*, pages 2329–2334. ACM, 2016.

[16] W. S. Ma, Y. Zheng. A large-scale dynamic taxi ridesharing service. In *The 29th International Conference on Data Engineering (ICDE'13)*, pages 410–421, New York, NY, 2013. IEEE.

[17] R. Witte and J. Witte. *Statistics, 10th Edition*. John Wiley and Sons, Incorporated, 2013.

[18] T. Xu, H. Zhu, X. Zhao, Q. Liu, H. Zhong, E. Chen, and H. Xiong. Taxi driving behavior analysis in latent vehicle-to-vehicle networks: A social influence perspective. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1285–1294. ACM, 2016.

[19] Y. Ge and H. Xiong and A. Tuzhilin and K. Xiao and M. Gruteser. An energy-efficient mobile recommender system. In *The the 16th International Conference on Knowledge Discovery and Data Mining*, pages 899–908, New York, NY, 2010. ACM.

[20] Y. Ge, C. Liu, H. Xiong, J. Chen. A Taxi Business Intelligence System. In *The 17th International Conference on Knowledge Discovery and Data Mining*, pages 735–738, New York, NY, 2011. ACM.

[21] 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*, pages 109–118, New York, NY, 2011. ACM.

[22] Z. Yuan, X. Zhou, and T. Yang. Hetero-convlstm: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 984–992. ACM, 2018.

[23] B. D. Ziebart. Modeling purposeful adaptive behavior with the principle of maximum causal entropy. 2010.

[24] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey. Maximum entropy inverse reinforcement learning. In *AAAI*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.