A Big Data Analysis on Urban Mobility: Case of Bangkok

SOOKSAN PANICHPAPIBOON, (Senior Member, IEEE) and KAVEPOL KHUNSRI
Faculty of Information Technology, King Mongkut’s Institute of Technology Ladkrabang, Bangkok 10520, Thailand
(e-mail: sooksan@alumni.cmu.edu, 60070003@kmitl.ac.th)
Corresponding author: Sooksan Panichpapiboon (e-mail: sooksan@alumni.cmu.edu)
This work was supported in part by the Faculty of Information Technology, King Mongkut’s Institute of Technology Ladkrabang.

ABSTRACT Designing an efficient on-demand mobility service requires comprehensive knowledge on the statistical characteristics of trips. In other words, it is critical to know how long passengers typically spend on a trip and how far they usually travel. Likewise, it is important to learn how much time a driver spends searching for passengers. This study presents a statistical analysis of taxi trips in Bangkok based on real traces of 5,853 taxis over the period of three months. Significant insights on trip volume, trip time, trip distance, and origin-destination distance are derived. In addition, the probability distributions of trip time, trip distance, and origin-destination distance are also characterized based on two goodness-of-fit tests. To our knowledge, this characterization is done for the first time for Bangkok taxi trips. It is shown that a lognormal distribution can best describe the empirical trip time distribution. On the other hand, a Weibull distribution can best describe the empirical trip distance distribution and the empirical origin-destination distance distribution. These distributions are essential to traffic simulation. Finally, the efficiency of the Bangkok taxi system is also quantified both at the system level and at the agent level.

INDEX TERMS Urban mobility, statistical modeling, data analytics, big data

I. INTRODUCTION
Understanding how people move in a city is critical to designing an effective transportation system. In-depth statistical knowledge on trip characteristics is required to build an efficient on-demand mobility service. In other words, it is necessary to know when people need a service, how long they typically spend on a trip, how far they usually travel, and how much time a driver typically spends searching for a passenger.

In this study, extending from our preliminary work [1], we perform statistical analysis on approximately eight million taxi trips in Bangkok, Thailand. Real traces of 5,853 taxis that operate in Bangkok and its nearby area during the period of three months from October 2019 to December 2019 are investigated. This allows us to thoroughly examine the spatio-temporal mobility pattern of Bangkok passengers and obtain significant insights into the characteristics of taxi trips in Bangkok. Moreover, it also allows us to evaluate the efficiency of the current Bangkok taxi system both at the system level and at the agent level.

In addition to understanding the characteristics of the current Bangkok taxi system, the statistical results presented in this study are beneficial to designing future mobility services such as on-demand self-driving taxis. Obviously, a service provider needs to understand the nature of taxi trips to devise an appropriate operation strategy. Moreover, the statistical results presented in this study are essential to traffic simulation. Indeed, a traffic simulator needs to know when and where to schedule the next taxi pickup realistically. It also needs to specify how long the trip should be and how far the destination should be from the origin. These require the statistical distributions derived from the empirical data. Finally, the efficiency of the current Bangkok taxi system evaluated in this study can be used as a benchmark for future mobility services. Basically, any future mobility service should be designed to achieve a higher efficiency level than that of the current system.

Although a few studies investigate the Bangkok taxis [2]–[5], none of them has characterized the statistical distributions of trip time, trip distance, and origin-destination (O-D) distance. In this paper, we characterize these distributions for the Bangkok taxi trips. Moreover, the data considered in this study are also more recent than those used in the existing studies. Unlike most existing works in the literature, which typically identify these distributions based on visual curve-fitting [6]–[12], we methodically characterize them based on
formal statistical tests.

The statistical analysis presented in this study covers many aspects of the taxi trips in Bangkok. The contributions of this work can be summarized as follows.

- We characterize the statistical distribution of trip time. It is shown that a lognormal distribution can best describe the trip time distribution. This applies to both the trips with passengers and those without passengers.
- We characterize the statistical distribution of trip distance. It is shown that a Weibull distribution can best describe the trip distance distribution. This applies to both the trips with passengers and those without passengers.
- We characterize the statistical distribution of O-D distance. It is shown that a Weibull distribution can best describe the O-D distance distribution. This applies to both the trips with passengers and those without passengers.
- At the system level, we quantify the overall efficiency of the Bangkok taxi system based on three months worth of positioning data. We demonstrate that the temporal efficiency of the Bangkok taxi system is only around 46.67%, and the spatial efficiency is only about 37.06%. This indicates that the efficiency of the Bangkok taxi system is poor.
- At the agent level, we quantify the performance of each taxi in terms of revenue-gain-rate. In addition, we also quantify the efficiency of individual taxis across different groups of earnings. We demonstrate that taxis in the high-earning group consistently have higher efficiency than those in the low-earning group. This suggests that taxis in the high-earning group do not simply earn more by working longer, but they are more efficient in driving and passenger searching.

The rest of this paper is organized as follows. In Section II, we briefly discuss related studies that rely on positioning data. Data and features extraction are discussed in Section III. The statistical analysis is presented in Section IV. Finally, we conclude this paper in Section V.

II. RELATED WORK

A number of studies that rely on positioning data of taxis exist. However, they can generally be divided into two main categories: 1) data analytics and 2) prediction. We briefly discuss works in these categories here.

1) Data Analytics

In data analytics, the taxi positioning data are typically used to investigate passengers’ mobility and drivers’ behaviors. Studies on passengers’ mobility usually involve analyzing the spatio-temporal characteristics of taxi trips such as trip time, trip distance, and O-D distance. In fact, many studies attempt to characterize the statistical distributions of these quantities [6]–[14]. However, only a visual observation based on curve-fitting is typically used in these studies to determine which statistical distribution can best fit the empirical data. A methodological verification by a statistical test is usually ignored.

Based on visual curve-fitting, the trip time distribution of taxi trips in Lisbon, Portugal is characterized with an exponential distribution [6]. In contrast, the trip time distribution of taxis in Xiamen, China, is characterized with a lognormal distribution in [7]. It is also reported that the trip time distribution of taxis in Shanghai, China, follows a lognormal distribution [8]. These characterizations are based on visual curve-fitting. A formal statistical test is not performed.

A variety of statistical distributions is used to characterize the trip distance distribution. A lognormal distribution is used to model the trip distance distribution of taxis in Harbin, China [9]. On the other hand, a gamma distribution is used to model the trip distance distribution of taxis in Lisbon, Portugal [6]. The gamma distribution is also used to model the trip distance distribution of taxis in Changsha, China [10]. In [11], the trip distance distribution of taxis in New York is characterized by a power-law distribution. Nonetheless, a statistical test is not employed in these studies. In [13], the statistical distribution of taxi trip distance in Chicago is investigated. A Kolmogorov-Smirnov (K-S) test is used to evaluate the goodness-of-fit. It is reported that the lognormal distribution can best describe the trip distance distribution.

The O-D distance is usually characterized with a power-law distribution. In [12], the authors investigate the taxi data in Beijing, China. The O-D distances of taxi trips that are smaller than 30 miles are modeled with a power-law distribution; whereas, the O-D distances of the larger trips are modeled with an exponential distribution. However, a goodness-of-fit test is not performed. In [14], the O-D distances of taxis trips in Shanghai, China, are characterized by a power-law distribution. A K-S test is used to evaluate the goodness-of-fit.

In addition to the characteristics of passengers’ mobility, many data-analytics works focus on taxi drivers’ behaviors. Typically, these studies involve analyzing passenger search strategies, identifying factors affecting cruising patterns, and examining route selection behaviors. In [15], the authors study the passenger search strategies of taxis in Shenzhen, China. A model based on Markov decision process (MDP) is proposed to imitate the passenger search strategies of the taxi drivers. In [16], three types of passenger search strategies of taxis in Hangzhou, China, are investigated. It is shown that local passenger hunting is usually more effective than local waiting and distant passenger hunting. In [17], factors affecting the cruising patterns of taxis in Shenzhen, China, are investigated. It is concluded that external factors (e.g., land use distribution, traffic condition, and local road network) have a stronger influence on each driver’s cruising pattern than internal factors (e.g., driver’s knowledge based on his past experience). Finally, in [18], the authors use the taxi data in Beijing, China, to analyze how a driver chooses a route to deliver a passenger. It is shown that most drivers usually choose a satisfactory route which may not necessarily be the optimal route (e.g., routes with shortest distance and shortest time).

It is interesting to note that Bangkok has not yet been
considered in most existing data-analytics studies. To our knowledge, only a few data-analytics studies have used positioning records of taxis in Bangkok [2], [3]. However, none of them has analyzed the statistical distributions of trip time, trip distance, and trip O-D distance. In addition, none of them has quantified the efficiency of the Bangkok taxi system. In [2], the impact of the new taxi fare rate on drivers and passengers is studied. The taxi service zones are analyzed and grouped based on the positioning records in [3].

2) Prediction

Most studies in this category typically focus on taxi demand prediction. An accurate demand prediction allows taxi drivers to cruise to the locations where they are more likely to find passengers. This also helps reduce the waiting time of a passenger. A variety of solutions is considered in the literature, including probabilistic models, time-series models, and machine learning models. In [19], the authors use the taxi data in San Francisco to build a probabilistic-based route recommendation model which aims at minimizing the mileage without passengers. In [20], the authors use the taxi data in Nanjing, China, to build a model for predicting the road cluster where the passengers would most likely be. The model is based on an extreme learning machine (ELM). It basically recommends top-k road clusters that a taxi is more likely to find passengers. In [21], the authors use the taxi data in Shanghai, China, to build a recommendation model that suggests where a taxi should look for passengers. The model is based on a deep convolutional neural network. In [22], the taxi data in Chengdu, China, are used to build a demand prediction model. Similarly, the taxi data in Porto, Portugal, Stockholm, Sweden, and Shanghai, China, are also used to build a demand prediction model in [23]. The model is based on an ensemble of autoregressive time-series models.

A few studies use Bangkok taxi data to build a demand prediction model. In [4], the authors propose a demand prediction model based on recurrent neural networks and XGBoost. In [5], the authors use k-means clustering to recommend a driver’s spot to wait for passengers.

The main focus of this paper is on data analytics rather than prediction. We analyze the spatio-temporal characteristics of taxi trips in Bangkok. Particularly, we perform a formal statistical test to characterize the distributions of trip time, trip distance, and O-D distance. This has not been done in the existing studies for Bangkok taxis.

III. DATA AND FEATURES EXTRACTION

In this section, we describe the data used in this study and the features extracted from these data.

A. DATA

The dataset used in this study is provided by courtesy of the Thai Intelligent Traffic Information Center (iTIC). It contains the global positioning system (GPS) records of Bangkok taxis during the period of three months from October 2019 to December 2019. The position of each taxi was collected roughly every one to three minutes. Essentially, the data have nine features as shown in Table 1. In addition to the primary positioning data such as latitude, longitude, and timestamp, the “for-hire” light status and the vehicle engine status are also provided. These features are beneficial for differentiating between trips with passengers and those without passengers. Basically, when the for-hire light status is ‘1’, it implies that the taxi is vacant. On the other hand, when the for-hire light status is ‘0’, it implies that the taxi is busy carrying a passenger.

In this study, we only focus on trips that occur in Bangkok and some parts of its nearby area. The study area is defined as the rectangular region shown in Fig. 1. The geolocations of the four corners of the study area are given in Table 2. In this study, we only concentrate on the period when each taxi is active (i.e., when the engine is running). Thus, all the records

### Table 1. Data description

| Feature       | Description                                      | Example               |
|---------------|--------------------------------------------------|-----------------------|
| VehicleID     | Anonymized unique vehicle ID                      | qdf+je...             |
| gpsvalid      | Logical value for valid GPS fix (1 = valid, 0 = invalid) | 1                     |
| lat           | Latitude                                         | 13.74184              |
| lon           | Longitude                                        | 100.49763             |
| timestamp     | GPS timestamp                                     | 2019-12-01 03:56:34   |
| speed         | Instantaneous speed (km/h)                       | 45                    |
| heading       | Vehicle heading direction (0-360 degrees, 0 = north) | 90                    |
| for_hire_light| Status of the “for hire” light (1 = light is on, 0 = light is off) | 1                     |
| engine_acc    | Status of the vehicle engine (1 = active, 0 = inactive) | 1                     |
with inactive engine status will be filtered out. In addition, all the records with invalid GPS status will also be filtered out as well. Finally, any records with the geolocations that are outside of the study area will be excluded. After this initial filtering process, we are left with the valid GPS records of the active taxis inside the study area.

### B. FEATURES EXTRACTION

After the initial filtering process described in Section III-A, the following features are extracted.

1) **Trip**: One of the most significant features to extract from the GPS records is trip. A taxi trip can be viewed as a time series of its consecutive GPS records. Basically, there are two types of trips. They are referred to as busy trips and vacant trips. A busy trip or a trip with passengers is defined as a time series of consecutive GPS records with inactive for-hire status. In other words, this is a sequence of consecutive GPS records with ‘0’ in the for-hire status. In contrast, a vacant trip or a trip without passengers is defined as a time series of consecutive GPS records with active for-hire status. In other words, this is a sequence of consecutive GPS records with ‘1’ in the for-hire status.

2) **Pickup and Drop-off Locations**: For a busy trip, we can estimate the pickup and drop-off locations. Basically, the beginning point of a busy trip is assumed to be the pickup location, and the endpoint of the trip is assumed to be the drop-off location. For example, suppose that a busy trip consists of $K$ consecutive GPS points, namely $P_1, P_2, \ldots, P_K$. Then, in this case, the estimated pickup location is $P_1$, and the estimated drop-off location is $P_K$.

3) **Trip Duration**: Trip duration is simply the difference between the timestamps at the start and at the end of the trip. Suppose that a trip $T$ consists of $K$ consecutive GPS points, namely $P_1, P_2, \ldots, P_K$. Then, the duration of trip $T$ can simply be calculated from the difference between the timestamp of $P_1$ and the timestamp of $P_K$.

4) **Trip Distance**: Trip distance is an aggregated sum of the distance between each pair of consecutive GPS points in the trip. Suppose that a trip $T$ consists of $K$ consecutive GPS points, namely $P_1, P_2, \ldots, P_K$. Let $D_{i,j}$ be a geographical distance between $P_i$ and $P_j$. Then, the total distance of trip $T$ is calculated as $D(T) = D_{1,2} + D_{2,3} + \ldots + D_{K-1,K}$. Note that this is not the Euclidean distance between the starting point and the endpoint of the trip.

5) **Origin-Destination Distance**: The origin-destination (O-D) distance is defined as the Euclidean distance between the starting point and the endpoint of the trip. If a trip $T$ consists of $K$ consecutive GPS points, namely $P_1, P_2, \ldots, P_K$, then the O-D distance of trip $T$ is simply the Euclidean distance between $P_1$ and $P_K$.

6) **Trip Fare**: A fare, in Thai Baht (THB), for each busy trip can also be estimated. Trip fare is calculated according to the Thai taxi fare rate imposed by the Department of Transportation. The total fare is a function of trip distance and congestion time. The distance-based fare rate is given in Table 3. Basically, the fare rate increases as the distance increases. The congestion time-based fare rate is 2 THB/minute for every minute that the taxi cannot travel faster than 6 km/h.

### IV. ANALYSIS

In this section, we present the analysis on the Bangkok taxi trips.

#### A. TRIP VOLUME

Trip volume reflects mobility demand. It helps us understand the transportation need in different periods of time. The total number of trips extracted from the dataset is shown in Table 4. There are approximately eight million trips in this analysis. Around 3.7 million of them are busy trips, and 4.3 million of them are vacant trips. The average daily trip volume on different days of the week is shown in Fig. 2. It can be observed that the average number of trips is approximately the same on each day of the week. On average, there are slightly over 40,000 trips each day. In addition, it can be

| Corner          | Latitude | Longitude |
|-----------------|----------|-----------|
| Upper left      | 13.94913 | 100.32750 |
| Upper right     | 13.94913 | 100.93963 |
| Lower left      | 13.49310 | 100.93963 |
| Lower right     | 13.49310 | 100.32750 |

#### TABLE 2. Four corners of the study area

#### TABLE 3. Bangkok Taxi Distance-based Fare Rate

| Distance (km) | Fare Rate (THB/km) |
|---------------|--------------------|
| ≤ 1           | 35.00              |
| 1-10          | 5.50               |
| 10-20         | 6.50               |
| 20-40         | 7.50               |
| 40-60         | 8.00               |
| 60-80         | 9.00               |
| > 80          | 10.50              |
TABLE 4. Number of Extracted Trips

| Trip Type | Amount     |
|-----------|------------|
| Busy      | 3,767,587  |
| Vacant    | 4,309,805  |
| Total     | 8,077,392  |

observed that the average number of trips on the weekdays and that on the weekends are not significantly different.

The average hourly trip volume in each hour of the day on different days of the week is shown in Fig. 3. It can be observed that the average hourly trip volume on all days of the week exhibits a common periodic pattern. Basically, on a given day, the average number of trips is small during the early morning hours. Among these hours, the lowest average number of trips is between 2 am and 3 am. The average number of trips rises in the morning rush hours and stays above 2,000 trips during the daytime. Finally, the average number of trips declines during the evening hours. However, in these evening hours (i.e., from 7 pm to 11 pm), the average number of trips on Saturday is noticeably larger than that of the others.

B. TRIP TIME

Trip time is an essential trait of a taxi trip. The busy trip time indicates the amount of time a passenger spends in a taxi whereas the vacant trip time indicates how long a taxi spends searching for a passenger. The probability density of the trip time extracted from this dataset is shown in Fig. 4. Both the busy trip time and the vacant trip time are compared. Generally, it can be observed that both the busy trip time distribution and the vacant trip time distribution are right-skewed. It implies that trips with a small duration occur more frequently than trips with a long duration. The mean duration of the busy trips and the mean duration of the vacant trips are approximately the same. Specifically, the mean duration of the busy trips is 24.14 minutes while the mean duration of the vacant trips is 24.02 minutes. However, the median duration of the vacant trips is slightly larger than that of the busy trips. Specifically, the median duration of the busy trips is 14.70 minutes whereas the median duration of the vacant trips is 15.72 minutes.

The natural question to ask next is the type of probability distributions that the empirical trip time distributions follow. In other words, what kind of probability distributions can be used to describe the empirical trip time distributions shown in Fig. 4? A common statistical method to test whether a hypothesized distribution can characterize an empirical distribution is the Kolmogorov-Smirnov test (K-S test) [26]. In the K-S test, the empirical distribution is initially assumed to follow a hypothesized distribution. Then, it is examined whether the hypothesis can be rejected. More formally, let $F(x)$ be the hypothesized distribution, and let $G(x)$ be the
empirical distribution. The null hypothesis and the alternative hypothesis for the K-S test are given as

\begin{align*}
H_0 & : G(x) = F(x) \\
H_1 & : G(x) \neq F(x)
\end{align*}

The null hypothesis will be rejected if the test result shows that the difference between the empirical distribution and the hypothesized distribution is statistically significant. Readers are referred to [26] for more detail on the K-S test.

Based on the observation that the empirical distributions are right-skewed, the following three types of distributions are selected as the hypothesized distributions in this study.

1) **Lognormal Distribution**: A lognormal distribution is a two-parameter probability model. Its cumulative distribution function (CDF) can be expressed as

\[
F_X(x) = \begin{cases} 
\frac{1}{2} + \frac{1}{2} \text{erf} \left( \frac{\ln x - \mu}{\sqrt{2} \sigma} \right), & x \geq 0 \\
0, & \text{otherwise}
\end{cases} \tag{1}
\]

where \( \text{erf}(y) \triangleq \frac{2}{\sqrt{\pi}} \int_0^y e^{-t^2} dt \) is the standard error function, \( \mu \in \mathbb{R} \) and \( \sigma^2 > 0 \).

2) **Gamma Distribution**: A gamma distribution is a two-parameter probability model. Its CDF can be expressed as

\[
F_X(x) = \begin{cases} 
\frac{\Gamma(\alpha, \beta x)}{\Gamma(\alpha)}, & x > 0 \\
0, & \text{otherwise}
\end{cases} \tag{2}
\]

where \( \alpha > 0 \) is the shape parameter, \( \beta > 0 \) is the rate parameter, \( \Gamma(s, x) \triangleq \int_x^\infty t^{s-1} e^{-u} du \) is the upper incomplete gamma function, and \( \gamma(s, x) \triangleq \int_0^x t^{s-1} e^{-u} du \) is the lower incomplete gamma function.

3) **Weibull Distribution**: A Weibull distribution is a two-parameter probability model. Its CDF can be expressed as

\[
F_X(x) = \begin{cases} 
1 - e^{-x/\lambda^k}, & x \geq 0 \\
0, & \text{otherwise}
\end{cases} \tag{3}
\]

where \( \lambda > 0 \) is the scale parameter, \( k > 0 \) is the shape parameter.

The following procedure is carried out to check whether these hypothesized distributions can be used to model the empirical trip time distributions. First, we create one thousand test sets, where each of these test sets contains two hundred random samples drawn from the empirical trip time data. Second, for each test set, we estimate the parameters of each hypothesized distribution that best fit the empirical samples based on the maximum likelihood estimation. In other words, the parameters \( \mu \) and \( \sigma \) of the lognormal distribution, the parameters \( \alpha \) and \( \beta \) of the gamma distribution, and the parameters \( \lambda \) and \( k \) of the Weibull distribution are estimated from the empirical samples in each test set. Finally, the K-S test is performed for each test set and each hypothesized distribution to check if the null hypothesis can be rejected. In this study, the null hypothesis will be rejected at the 5% significance level.

### Table 5. Percentage of the test sets that pass the K-S test and the A-D test in trip time characterization

| Test | Trip Type | Hypothesized Distribution |
|------|-----------|---------------------------|
|      |           | Lognormal | Gamma | Weibull |
| K-S  | Busy      | 98.6%     | 86.3% | 97.9%   |
| K-S  | Vacant    | 98.5%     | 78.4% | 88.6%   |
| A-D  | Busy      | 99.9%     | 89.7% | 96.6%   |
| A-D  | Vacant    | 99.3%     | 80.8% | 78.6%   |

In addition, in order to cross-check the results obtained from the K-S test, the Anderson-Darling (A-D) goodness-of-fit test is also performed. Similar to the K-S test, the A-D test evaluates if a set of random samples could have come from a hypothesized distribution. Readers are referred to [27] for more detail on the A-D test. The same experimental procedure as described in the previous paragraph is repeated for the A-D test.

The percentage of the test sets that pass the K-S test (i.e., those for which the null hypothesis cannot be rejected) for each type of hypothesized distributions is shown in Table 5. It can be observed that the lognormal distribution can best describe the busy trip time distribution. In fact, it can describe 98.6% of the test sets. This literally means that, out of 1,000 test sets, 986 of them can be characterized by the lognormal distribution. The Weibull distribution can also model the busy trip time distribution very well. It can describe 97.9% of the test sets. The gamma distribution, however, is not as effective in modeling the busy trip time distribution. It can only describe 86.3% of the test sets. In the case of the vacant trip time distribution, the only distribution that can best describe it is the lognormal distribution. The lognormal distribution can model 98.5% of the test sets while the other two distributions are much less effective.

A similar trend is also observed with the A-D test results shown in Table 5. In the case of the busy trips, the lognormal distribution and the Weibull distribution can model the trip time distribution exceptionally well. In fact, 99.9% of the test sets pass the A-D test when the hypothesized distribution is lognormal. In the case of the vacant trips, however, the only distribution that can best describe the trip time distribution is the lognormal distribution. The A-D test results are coherent with the K-S test results.

Next, we investigate if these hypothesized distributions are able to model the trip time distribution in each hour of the day. In addition, we also investigate if there is any difference between weekday trips and weekend trips. The same experimental procedure as described earlier is repeated for each hour of the day. The percentage of the test sets that pass the K-S test in each hour of the day is shown in Fig. 5. In the case of the busy trips (shown in the top row), the lognormal distribution can model the trip time distribution remarkably well in all hours of the day. Moreover, it can model the trip time distribution both on the weekday and on the weekend.
The Weibull distribution is also able to model the trip time distribution remarkably well on almost all hours of the day, both on the weekday and on the weekend. However, there are a few early morning hours on the weekend that the Weibull distribution becomes less effective. In the case of the vacant trips (shown in the bottom row), the lognormal distribution is the only distribution that can model the trip time distribution extremely well. Similar behaviors are also observed with the A-D test results shown in Fig. 6.

In summary, the lognormal distribution is the only distribution that can model both the busy trip time distribution and the vacant trip time distribution exceptionally well. Both the K-S test results and the A-D test results confirm this.

C. TRIP DISTANCE

Trip distance is also an important trait of a taxi trip. The busy trip distance shows the characteristics of the passenger delivery route while the vacant trip distance indicates how far a taxi needs to drive to search for a passenger. The probability density of the trip distance is shown in Fig. 7. Both the busy trip distance and the vacant trip distance are compared. Similar to the trip time distribution, both the busy trip distance distribution and the vacant trip distance distribution are right-skewed. This means that short-distance trips occur more frequently than long-distance trips. The mean distance of the busy trips is 6.43 km while the mean distance of the vacant trips is 9.55 km. This indicates that the distance incurred on passengers searching is typically larger than the distance incurred on passengers delivering. The median distance of the busy trips is 3.88 km while the median distance of the vacant trips is 6.15 km. These numbers clearly suggest that the current taxi system is inefficient. The taxis typically need to drive longer to search for a passenger than actually deliver one.

Next, we characterize the empirical trip distance distributions. The same statistical procedure as described in Section IV-B is also performed on the empirical trip distance data. The percentage of the test sets that pass the K-S test for each type of hypothesized distributions is shown in Table 6. It can be observed that the Weibull distribution and the gamma distribution can model the busy trip distance distribution remarkably well. The Weibull distribution can describe 99.7% of the test sets while the gamma distribution can describe 98.9% of the test sets. The lognormal distribution is fair in
D. ORIGIN-DESTINATION DISTANCE

Origin-Destination distance is the Euclidean distance between the starting point and the endpoint of a trip. The statistical knowledge on the O-D distance distribution is critical to creating a realistic O-D pair in a traffic simulation. The probability density of the O-D distance is shown in Fig. 10. Both the busy trip O-D distance and the vacant trip O-D distance are compared. Similar to the trip distance distribution, both the busy trip O-D distance distribution and the vacant trip O-D distance distribution are right-skewed. The mean O-D distance of the busy trips is 4 km. It implies that, on average, the passengers typically use the taxis to get to the destinations within the 4 km radius. The mean O-D distance of the vacant trips is 6.36 km. It implies that, on average, the taxis typically search for the passengers within the 6.36 km radius. The median O-D distance of the busy trips is 2.41 km while the median O-D distance of the vacant trips is 4.29 km.
TABLE 7. Percentage of the test sets that pass the K-S test and the A-D test in O-D distance characterization

| Test | Trip Type | Hypothesized Distribution | Lognormal | Gamma | Weibull |
|------|-----------|---------------------------|-----------|-------|--------|
| K-S  | Busy      | 96.8%                     | 97.2%     | 99.7% |
| K-S  | Vacant    | 14.9%                     | 88.4%     | 97.6% |
| A-D  | Busy      | 98.6%                     | 98.6%     | 99.9% |
| A-D  | Vacant    | 12.4%                     | 86.3%     | 93.3% |

Next, we characterize the empirical O-D distance distributions. The same statistical procedure as described in Section IV-B is also performed on the empirical O-D distance data. The percentage of the test sets that pass the K-S test for each type of hypothesized distributions is shown in Table 7. In the case of the busy trips, all three types of hypothesized distributions are able to model the O-D distance distribution extremely well although the Weibull distribution is the best. The Weibull distribution is able to model 99.7% of the test sets. In the case of the vacant trips, in contrast, only the Weibull distribution can model the O-D distance distribution effectively. The Weibull distribution can model 97.6% of the test sets.

The percentage of the test sets that pass the A-D test for each type of hypothesized distributions is also shown in Table 7. A similar trend as noted from the K-S test results can also be observed. In the case of the busy trips, all three types of hypothesized distribution can model the trip O-D distance distribution extremely well. In the case of the vacant trips, the O-D distance can best be described by the Weibull distribution.

The percentage of the test sets that pass the K-S test in each hour of the day is shown in Fig. 11. In the case of the busy trips, all three types of hypothesized distributions can model the O-D distance distribution exceptionally well in all hours of the day, both on the weekday and on the weekend. In the case of the vacant trips, the Weibull distribution can model the O-D distance distribution effectively. The percentage of the test sets that pass the A-D test in each hour of the day in trip O-D distance characterization.

Both the daytime (i.e., 6 am to 6 pm) locations and the night time (i.e., 6 pm to 6 am) locations are illustrated. Generally, the density of the pickups concentrates around the city center. The density decreases as the distance from the center increases. During the daytime, there are four distinct bright spots on the density map. These spots are the city center area, the Chatuchak area, the Don Mueang International Airport (DMK), and the Bangkok International Airport (BKK). The city center area is where most business activities take place. Thus, a high intensity of pickups is expected. The Chatuchak area is where many transportation terminals (e.g., bus, subway, and train) are located, so this is probably the main reason why a lot of pickups are observed in this area. During the night time, the density map generally looks similar to that of the daytime. However, the density of the pickups during the night time is highest around the Ploen Chit area where many of the shopping malls and hotels are located.

E. PICKUP AND DROP-OFF LOCATIONS

Pickup and drop-off locations can help us visualize where people travel. A density map of the pickup locations is shown in Fig. 13. The density is normalized so that it ranges between 0 and 1, where 1 represents the highest density.
FIGURE 13. Density map of pickup locations. (This map is created from the open data provided by OpenStreetMap [24] under the Open Database Licence [25].)

FIGURE 14. Density map of drop-off locations. (This map is created from the open data provided by OpenStreetMap [24] under the Open Database Licence [25].)
A density map of the drop-off locations is shown in Fig. 14. Similar to what observed in the case of the pickups, the density of the drop-offs generally concentrates around the city center. It decreases as the distance from the center increases. During the daytime, there are two distinct bright spots. They are the city center area and the Chatuchak area. During the night time, the density of the drop-offs is highest around the Ploen Chit area, which is similar to what observed in the case of the pickups at night.

F. SYSTEM EFFICIENCY

In addition to the statistical characteristics of the taxi trips discussed in the previous sections, in this section, we evaluate the efficiency of the Bangkok taxi system. Particularly, the system efficiency is measured by the following two metrics.

1) Temporal efficiency: Temporal efficiency is defined as the ratio between the total busy time contributed by all taxis and the total operating time of all taxis. It basically measures the proportion of time that the system is busy. Formally, the temporal efficiency can be expressed as

\[
\eta_T = \frac{\sum_{i=1}^{N} \tau_{B_i}}{\sum_{i=1}^{N} (\tau_{B_i} + \tau_{V_i})}
\]

where \( \tau_{B_i} \) is the amount of time that taxi \( i \) is busy, \( \tau_{V_i} \) is the amount of time that taxi \( i \) is vacant, and \( N \) is the total number of taxis in the system.

2) Spatial efficiency: Spatial efficiency is defined as the ratio between the total busy distance contributed by all taxis and the total operating distance of all taxis. Formally, the spatial efficiency can be expressed as

\[
\eta_S = \frac{\sum_{i=1}^{N} d_{B_i}}{\sum_{i=1}^{N} (d_{B_i} + d_{V_i})}
\]

where \( d_{B_i} \) is the amount of distance that taxi \( i \) is busy, \( d_{V_i} \) is the amount of distance that taxi \( i \) is vacant, and \( N \) is the total number of taxis in the system.

The efficiency of the Bangkok taxi system is shown in Table 8. It is clear that the Bangkok taxi system is highly inefficient. The temporal efficiency is only around 46.76%, which implies that the taxis spend more time searching for passengers than actually carrying them. The spatial efficiency is worse. The total distance traveled by the taxis while carrying the passengers is only 37.06% of the total operating distance. The temporal efficiency and the spatial efficiency reported here are critical to the future development of on-demand mobility services. They are the benchmark that future services need to take into consideration. Obviously, any future services should be designed to have a higher efficiency level.

G. AGENT EFFICIENCY

In this section, we evaluate the efficiency of each individual taxi. In this analysis, a taxi is viewed as a revenue-generating agent in the system. Since a taxi could be driven by more than one driver, the efficiency computed in this analysis is not intended to reflect the performance of an individual driver. Rather, it reflects how well each individual taxi (i.e., machine) contributes as a revenue-generating asset in the system. The efficiency of an individual taxi is measured in terms of revenue-gain-rate. The revenue-gain-rate is defined as the amount of revenue (i.e., fare) generated by a taxi over its operating time. It is measured in THB/hour, and it can be expressed as

\[
\gamma_i = \frac{r_i}{\tau_{B_i} + \tau_{V_i}}
\]

where \( r_i \) is the revenue generated by taxi \( i \), \( \tau_{B_i} \) is the amount of time that taxi \( i \) is busy, and \( \tau_{V_i} \) is the amount of time that taxi \( i \) is vacant.

Since we are assessing how much revenue each taxi gains, it is logical to consider only the taxis that operate regularly. Therefore, to evaluate the revenue-gain-rate, we only consider the taxis that make at least 100 busy trips during the three-month study period. This is roughly equivalent to making at least one busy trip a day, on average. The probability density of the revenue-gain-rate is shown in Fig. 15. The distribution looks symmetrical. The mean revenue-gain-rate is 102.45 THB/hour, and the median revenue-gain-rate is 103.29 THB/hour. The minimum wage in Thailand is 300 THB/day or around 37.5 THB/hour. Thus, the mean revenue-gain-rate is approximately 2.7 times higher than the minimum wage rate.

Next, we divide the taxis into four quantiles based on the total revenue they generate during the three-month study period. The efficiency of an individual taxi is measured in terms of revenue-gain-rate. The revenue-gain-rate is defined as the amount of revenue (i.e., fare) generated by a taxi over its operating time. It is measured in THB/hour, and it can be expressed as

\[
\gamma_i = \frac{r_i}{\tau_{B_i} + \tau_{V_i}}
\]

where \( r_i \) is the revenue generated by taxi \( i \), \( \tau_{B_i} \) is the amount of time that taxi \( i \) is busy, and \( \tau_{V_i} \) is the amount of time that taxi \( i \) is vacant.

Since we are assessing how much revenue each taxi gains, it is logical to consider only the taxis that operate regularly. Therefore, to evaluate the revenue-gain-rate, we only consider the taxis that make at least 100 busy trips during the three-month study period. This is roughly equivalent to making at least one busy trip a day, on average. The probability density of the revenue-gain-rate is shown in Fig. 15. The distribution looks symmetrical. The mean revenue-gain-rate is 102.45 THB/hour, and the median revenue-gain-rate is 103.29 THB/hour. The minimum wage in Thailand is 300 THB/day or around 37.5 THB/hour. Thus, the mean revenue-gain-rate is approximately 2.7 times higher than the minimum wage rate.
period. The first quantile represents the bottom 25% of taxis with the lowest revenue whereas the fourth quantile represents the top 25% of the taxis with the highest revenue. The revenue-gain-rate of the taxis that belong in each quantile is shown in Fig. 16. It can be observed that the taxis in the fourth quantile generally have the highest revenue-gain-rate. This suggests that the taxis in this group do not earn more revenue by simply operating longer. On the contrary, these taxis are more efficient in searching for passengers and generating revenues.

V. CONCLUSIONS

In this study, we perform statistical analysis on the Bangkok taxi trips, which are extracted from the real GPS records of taxis in the period between October 2019 and December 2019. We conclude this study with the following insights.

- In terms of trip volume, the average number of trips on each day of the week is approximately the same. In addition, there is no significant difference between the average number of trips on the weekday and that on the weekend.
- In terms of trip time, Bangkok passengers spend, on average, around 24.14 minutes on each taxi trip. Similarly, the taxis spend, on average, around 24.02 minutes searching for a passenger. Based on the K-S test and the A-D test, the probability distribution that can best describe both the busy trip time distribution and the vacant trip time distribution is the lognormal distribution. This agrees with [7] and [8], where the busy trip time distribution is characterized with a lognormal distribution based on visual curve-fitting.
- In terms of trip distance, the driving distance between the pickup location and the drop-off location is around 6.43 km, on average. However, the average distance that a taxi drives to search for passengers is around 9.55 km. Based on the K-S test and the A-D test, the probability distribution that can best describe both

![FIGURE 16. Revenue-gain-rate of taxis in each revenue quantile.](image-url)
I. Martínez and W.-L. Jin, “On time-dependent trip distance distribution,” IEEE Commun. Mag., vol. 56, no. 3, pp. 142–149, Mar. 2018.

S. Zhang, J. Tang, H. Wang, Y. Wang, and S. An, “Revealing intra-urban travel patterns and service ranges from taxi trajectories,” J. Transp. Geography, vol. 61, pp. 72–86, May 2017.

H. Rong, Z. Wang, H. Zheng, C. Hu, L. Peng, Z. Ai, and A. K. Sangai, “Mining efficient taxi operation strategies from large scale geo-location data,” IEEE Access, vol. 5, pp. 25 623 – 25 634, July 2017.

X. Qian, X. Zhao, and S. V. Ukkusuri, “Characterizing urban dynamics using large scale taxi cab data,” in Engineering and Applied Sciences Optimization, ser. Computational Methods in Applied Sciences, N. Lagaros and M. Papadrakakis, Eds. Springer, 2015, vol. 38, pp. 17–32.

H. Cai, X. Zhan, J. Zhu, X. Jia, A. S. Chiu, and M. Xu, “Understanding taxi travel patterns,” Physica A: Statistical Mechanics and Its Appl., vol. 457, pp. 590–597, Sept. 2016.

I. Martínez and W.-L. Jin, “On time-dependent trip distance distribution with for-hire vehicle trips in chicago,” Transp. Res. Rec., vol. 2675, no. 11, pp. 915–934, 2021.

Y. Liu, C. Kang, S. Gao, Y. Xiao, and Y. Tian, “Understanding intra-urban trip patterns from taxi trajectory data,” J. Geogr. Syst., vol. 14, p. 463–483, Mar. 2012.

X. Zhang, Y. Li, X. Zhou, and J. Luo, “cGAIL: Conditional generative adversarial imitation learning—an application in taxi drivers’ strategy learning,” IEEE Trans. Big Data, 2020, in press.

D. Zhang, L. Sun, B. Li, C. Chen, G. Pan, S. Li, and Z. Wu, “Understanding taxi service strategies from taxi GPS traces,” IEEE Trans. Intell. Transp. Syst., vol. 16, no. 1, pp. 123–135, Feb. 2015.

F. Zong, T. Wu, and H. Jia, “Taxi drivers’ cruising patterns—insights from taxi GPS traces,” IEEE Trans. Intell. Transp. Syst., vol. 20, no. 2, pp. 571–582, Feb. 2019.

L. Li, S. Wang, and F. Wang, “An analysis of taxi driver’s route choice behavior using the trace records,” IEEE Trans. Computat. Social Syst., vol. 5, no. 2, pp. 576–582, June 2018.

Z. Luo, H. Lv, F. Fang, Y. Zhao, Y. Liu, X. Xiang, and X. Yuan, “Dynamic taxi service planning by minimizing cruising distance without passengers,” IEEE Access, pp. 70 005–70 016, Nov. 2018.

R. Wang, C.-Y. Chow, Y. Lyu, V. C. S. Lee, S. Kwong, Y. Li, and J. Zeng, “TaxiRec: recommending road clusters to taxi drivers using ranking-based extreme learning machines,” IEEE Trans. Knowl. Data Eng., vol. 30, no. 3, pp. 585–598, Mar. 2018.

Z. Huang, J. Tang, G. Shan, J. Ni, Y. Chen, and C. Wang, “An efficient passenger-hunting recommendation framework with multitask deep learning,” IEEE Internet of Things J., vol. 6, no. 5, pp. 7713–7721, Oct. 2019.

K. Niu, C. Cheng, J. Chang, H. Zhang, and T. Zhou, “Real-time taxi-passer prediction with L-CNN,” IEEE Trans. Veh. Technol., vol. 68, no. 5, pp. 4122–4129, May 2019.

A. Saadallah, L. Moreira-Matias, R. Sousa, J. Khiari, E. Jenelius, and J. Gama, “BRIGHT—drift-aware demand predictions for taxi networks,” IEEE Trans. Knowl. Data Eng., vol. 32, no. 2, pp. 234–245, Feb. 2020.

OpenStreetMap,” https://www.openstreetmap.org, (accessed June 5, 2021).

“Open Data Commons,” https://opendatacommons.org, (accessed June 5, 2021).

D. Sheskin, Handbook of Parametric and Nonparametric Statistical Procedures, 4th ed. Chapman & Hall, 2007.

T. W. Anderson and D. A. Darling, “A test of goodness of fit,” J. Amer. Statistical Assoc., vol. 49, no. 268, pp. 765–769, Dec. 1954.

***

SOOKSAN PANICHPAPIBOON (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Carnegie Mellon University, Pittsburgh, PA, USA, in 2000, 2002, and 2006, respectively, all in electrical and computer engineering. He is currently a Full Professor with the Faculty of Information Technology, King Mongkut’s Institute of Technology Ladkrabang, Bangkok, Thailand. His current research interests include intelligent transportation systems, vehicular networks, mobile sensors, and performance modeling.

KAVEPOL KHUNSRI is currently a student in an integrated bachelor’s and master’s degree program in information technology at the Faculty of Information Technology, King Mongkut’s Institute of Technology Ladkrabang, Bangkok, Thailand. His research interests are big data analytics and machine learning. In 2021, he won the best paper award from the 18th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications, and Information Technol-