IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 April 2024, accepted 21 April 2024, date of publication 29 April 2024, date of current version 6 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3394497

RESEARCH ARTICLE

Characterizing the Distributions of Taxi Demand: Is Poisson the Right Model?

SOOKSAN PANICHPAPIBOON, (Senior Member, IEEE), AND KAVEPOL KHUNSRI
School of Information Technology, King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand

Corresponding author: Sooksan Panichpapiboon (sooksan@alumni.cmu.edu)

This work was supported in part by the School of Information Technology, King Mongkut's Institute of Technology Ladkrabang under Grant 2566-02-06-003.

ABSTRACT Statistical distribution of taxi demand is essential for modeling the dynamics of taxi services. An accurate demand prediction not only helps the drivers lessen their searching time but also helps the passengers shorten their waiting time. Moreover, the temporal distribution of taxi demand is a critical component in traffic simulation. Obviously, a simulator needs to know how many new taxi pickup events to schedule after the others. In most studies, a poisson distribution is often used to model the temporal distribution of taxi pickups. However, this assumption has mostly been used without validation with empirical data. Therefore, it is unclear whether such an assumption is appropriate for modeling the statistical distribution of taxi demand. In this study, we characterize the temporal distribution of taxi pickups based on real taxi trip data from Bangkok, Thailand, and Chicago, IL, USA. It is shown that, in most cases, the poisson distribution is not suitable for modeling the temporal distribution of taxi pickups. On the contrary, this study demonstrates that a geometric distribution is more appropriate in modeling the temporal distribution of taxi pickups. To our knowledge, this has not been discovered in any prior studies.

INDEX TERMS Taxi demand, statistical modeling, statistical analysis, poisson distribution, intelligent transportation systems.

I. INTRODUCTION

Knowledge of the temporal distribution of taxi demand is vital for improving the efficiency of taxi services. Understanding the underlying statistical distributions of demand is absolutely crucial to demand prediction. An accurate demand prediction allows vacant taxis to cruise to the areas where they are more likely to find passengers. This not only helps the drivers lessen their searching time but also helps the passengers shorten their waiting time. In addition, the temporal distribution of taxi demand is a critical component in traffic simulation. Obviously, a simulator needs to know how many new taxi pickup events to schedule after the others. These are determined by the temporal distribution of taxi pickups.

In most studies, it is frequently assumed that the temporal distribution of the number of taxi pickups follows a poisson distribution [1], [2], [3], [4], [5], [6]. Nonetheless, this assumption has mostly been used without validation with empirical data. One of the main reasons why a poisson distribution is often assumed is that many of its properties can help simplify an analysis. First, when an arrival process (e.g., arrivals of pickup requests) is poisson, it automatically implies that the interarrival time (i.e., the time between consecutive arrivals) follows an exponential distribution. This tremendously helps save the troubles of characterizing the interarrival time distribution, which could be difficult to derive had the other types of arrival process been assumed. Second, by assuming that both the arrivals of pickup requests and the arrivals of vacant taxis (i.e., services) are poisson, the system's dynamics can be modeled with a Markovian queue. The Markovian queues have been thoroughly analyzed; therefore, the known results can readily be applied. Using other arrival processes would not allow the system to be conveniently modeled as a Markovian queue. However, any analysis is meaningless unless the underlying

The associate editor coordinating the review of this manuscript and approving it for publication was Ivan Wang-Hei Ho.

assumption is valid. In this paper, we will verify whether a poisson distribution is appropriate for modeling the temporal distribution of the number of pickups. If it is not, we will suggest a suitable statistical distribution that can effectively model the number of pickups.

In this study, the temporal distribution of the number of pickups is characterized based on real taxi trips in Bangkok, Thailand, and Chicago, IL, USA. A statistical procedure is performed to test whether any particular statistical distribution can model the temporal distribution of the number of pickups. Three types of discrete probability distributions are investigated. They are poisson distribution, negative binomial distribution, and geometric distribution. Since the poisson distribution is frequently assumed in most studies without validation with empirical data, we will examine it here. The negative binomial distribution was used to model the temporal distribution of the number of pickups in [7]; therefore, it is also included for investigation here. Finally, to our knowledge, the geometric distribution has not been used to model the temporal distribution of the number of pickups in any existing studies. It is newly introduced here to characterize the temporal variation of the number of pickups.

The contributions of this work can be summarized as follows.

1) With two independent datasets, we characterize the temporal distribution of the number of pickups. The characterized distribution is crucial to future development in the theoretical analysis and simulation of taxi demand. It provides a fundamental basis for researchers to formulate new theories and build models upon.
2) We verify with real data that, in most cases, the poisson distribution is ineffective in modeling the temporal distribution of the number of pickups. In fact, we demonstrate that the poisson distribution can only model the temporal distribution in the scenario where the mean number of pickups is extremely small. Although it might be well-established within traffic flow modeling that the poisson distribution is appropriate for modeling vehicle arrivals in light traffic scenarios, this is yet to be established within the context of taxi demand. The finding that the poisson distribution is only appropriate for the scenario where the mean number of pickups is extremely small is substantiated for the first time in this study.
3) Finally, we discover a new distribution that is more appropriate than the poisson distribution in modeling the temporal distribution of the number of pickups. While the effectiveness of the poisson distribution is only limited to the scenario where the mean number of pickups is extremely small, we demonstrate that the geometric distribution is effective in broader scenarios with a higher mean number of pickups. The geometric distribution has never been used to model the distribution of taxi pickups in any existing work. It is introduced for the first time in this study.

The rest of this paper is organized as follows. In Section II, we briefly discuss related work. The methodology used in this study is presented in Section III. Results are discussed in Section IV. Finally, we conclude this study in Section V.

II. RELATED WORK

The statistical distributions of demand and supply are crucial to studies on mobility services such as traditional taxis and on-demand ride-sharing. A number of studies on demand and supply in mobility services exist. These studies typically involve predicting short-term demand and supply, evaluating the quality of service, and rebalancing demand and supply in the system. All of them have to make critical assumptions on the temporal distributions of demand and supply. In order to stress the importance of these distributions, we briefly review related works in which the distributions of demand and supply are required.

Being able to predict demand and supply in a mobility service system is beneficial to both the customers and the service provider. For example, in the case of a traditional taxi system, an accurate demand prediction allows the taxis to cruise to the regions where they are more likely to get passengers. Similarly, an accurate supply prediction allows the passengers to know when and where they are more likely to find vacant taxis. In most of the studies on demand and supply prediction, it is usually assumed that the demand (e.g., the number of passenger arrivals or the number of service requests), as well as the supply (e.g., the number of arrivals of vacant taxis), follows a poisson distribution [1], [2], [3], [4], [5], [6]. In addition, to cope with the time-varying nature of demand and supply, a non-homogeneous poisson distribution is typically employed. In [1] and [2], a combination of a non-homogeneous poisson model and an autoregressive integrated moving average (ARIMA) model was used in predicting passenger demand in a given region. In [3], the impact of spatial resolution on the demand prediction of an autonomous taxi service was investigated. It was assumed that the number of service requests followed a non-homogeneous poisson distribution. In [4], the authors used non-homogenous poisson distributions to model and predict the demand and supply of a traditional taxi system in Munich, Germany. However, only the daily average number of pickups could be predicted instead of the actual number of pickups. The empirical distribution was not characterized. In [5], the authors proposed a recommendation system for taxis and passengers. It was assumed that the number of arrivals of vacant taxis on a given road segment followed a non-homogeneous poisson distribution. In [6], a short-term demand prediction model for a bus system was proposed. It was assumed that the number of passenger arrivals at a bus stop followed a non-homogeneous poisson distribution.

Many studies concentrate on modeling the dynamics of mobility service systems. In these studies, queuing models and simulation models are often used to describe and mimic the stochastic interaction between elements of the system. In all of these models, it is assumed that the number of passenger arrivals or service requests follows a poisson distribution [8], [9], [10], [11], [12], [13], [14]. In [8], a taxi stand operation was described by a queuing model. It was assumed that the number of passenger arrivals and the number of taxi arrivals followed poisson distributions. In [9], a queuing model was proposed to capture the dynamics between the demand and the supply of a street-hail taxi system at the street segment level. The arrivals of passengers and the arrivals of vacant taxis were modeled by poisson distributions. The poisson arrival assumption was also validated on the NYC taxi data during the morning peak hour from 8 am to 9 am on Tuesdays to Thursdays. In [10], the authors also proposed a queuing model, which assumed that the number of passenger arrivals and the number of taxi arrivals could be described by poisson distributions. A hypothesis test was performed on the NYC taxi data to verify this assumption. It was shown that the poisson assumption only limitedly held in the one-hour peak period from 6 pm to 7 pm and in the one-hour off-peak period from 10 am to 11 am. In [11] and [12], the waiting time of the passengers was estimated. The number of passenger arrivals and the number of taxi arrivals were modeled using poisson distributions. In addition to the traditional taxi services, other mobility services have also been considered. In [13], a vehicle rental system was modeled with a queuing network, where the number of customer arrivals was assumed to follow a poisson distribution. In [14], the number of customer arrivals in an air taxi network was also assumed to follow a poisson distribution.

Rebalancing of demand and supply in a mobility service system has also been investigated in many studies [15], [16], [17], [18]. In [15], the authors proposed a solution to re-position bikes among the sharing stations. It was assumed that the number of customer arrivals at the sharing stations followed a poisson distribution. Similarly, in [16], an algorithm to assign idle vehicles in a ride-sharing service to the regions that required rebalancing was proposed. The number of request arrivals in each region was assumed to follow a non-homogeneous poisson distribution. In [17], the authors proposed a queuing network to model a mobility service system where autonomous robots could rebalance themselves to maintain the desired quality of service. The number of customer arrivals was assumed to follow a poisson distribution. In [18], the authors proposed an algorithm to redistribute empty autonomous taxis so that the waiting time of the passengers was minimized. It was assumed that the number of passenger arrivals followed a poisson distribution.

As observed in the literature, the poisson arrival assumption has been used extensively without validation with real data in most of the works. Only a few studies directly investigated the temporal distribution of demand in mobility service systems. In [19], the authors used samples from the NYC taxi data to evaluate how different variants of poisson models can represent the temporal variation of the pickups. However, these variants of poisson models were not compared with other types of distributions. In [7], the authors compared the ability of the negative binomial distribution and the ability of the poisson distribution to model the temporal variation of the pickups. The evaluation was done on the NYC taxi data. It was shown that the negative binomial distribution was more appropriate for modeling the pickup count data.

In this paper, we characterize the temporal distribution of taxi pickups based on real taxi trips in Bangkok and Chicago. In fact, we demonstrate that, in most scenarios, the geometric distribution can capture the temporal variation of the pickups more effectively than the poisson distribution and the negative binomial distribution. This new discovery has not been reported in any existing studies.

III. METHODOLOGY

In this section, we provide details on the datasets and the method used to analyze the temporal distribution of the number of pickups.

A. DATA

In this study, we analyze the taxi trips in December 2019 from two independent datasets. One of them is from Bangkok, and the other is from Chicago. In the following sections, we provide details on the characteristics of these datasets.

1) BANGKOK TAXI DATA

The first dataset used in this study is publicly provided by the Thai Intelligent Traffic Information Center (iTIC) [20]. It contains real global positioning system (GPS) records of Bangkok taxis. The data were collected from each taxi roughly every one to three minutes. Besides the primary positioning data, such as latitude, longitude, and timestamp, the "for-hire" light status and the vehicle engine status were also collected. The status of the for-hire light is a logical variable, where 0 indicates that the light is off and 1 indicates that the light is on. Similarly, the vehicle engine status is also a logical variable, where 0 indicates that the engine is off and 1 indicates that the engine is on. These features are essential for differentiating between busy trips and vacant trips. Basically, when the status of the for-hire light is 1, it can be implied that the taxi is vacant. In contrast, when the status of the for-hire light is 0, it can be implied that the taxi is carrying a passenger.

The study area in the Bangkok dataset is defined as the rectangular region shown in Fig. 1. The geolocations of the four corners of the study area are given in Table 1. Any data with geolocations outside of the study area will be filtered out. In addition, we only concentrate on the period when each taxi is active (i.e., when the engine is running). Thus, all the data with inactive engine status will be filtered out as well. Finally, after this filtering process, we are left with the records of taxis that are active inside the study area.

Next, we need to identify where the pickup locations are. A pickup location is essentially the starting point of a *trip*. A taxi trip can be viewed as a sequence of consecutive GPS records with the same for-hire status. Basically, there are

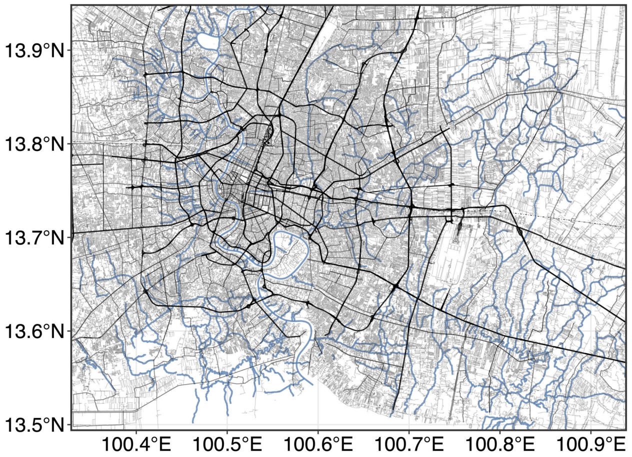

FIGURE 1. The study area considered in the Bangkok dataset is shown inside the rectangular bounding box. (This map is created from the open data provided by OpenStreetMap [21] under the Open Database Licence [22].)

TABLE 1. Four corners of the study area in the Bangkok dataset.

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

two types of trips: 1) busy trip and 2) vacant trip. A busy trip or a trip with passengers is defined as a sequence of consecutive GPS records with inactive for-hire status (i.e., 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 sequence of consecutive GPS records with active for-hire status (i.e., a sequence of consecutive GPS records with ‘1’ in the for-hire status). Since we are merely interested in the distribution of pickups, only the busy trips will be considered. 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. Suppose that a busy trip consists of five consecutive GPS points, for example, $P_1 \rightarrow P_2 \rightarrow P_3 \rightarrow P_4 \rightarrow P_5$. Then, in this case, the estimated pickup location is P_1, and the estimated drop-off location is P_5.

An additional cleanup is performed in order to filter out some possibly erroneous data. First, trips with abnormally long and abnormally short duration will be excluded. In this study, trips that are longer than or equal to 4 hours and trips that are shorter than or equal to 1 minute are considered abnormal. Second, trips with unusually large distances will also be filtered out. The maximum driving distance from one corner of the study area to its diagonally opposite corner is around 120 km. Thus, trips that are larger than 120 km will be considered abnormal. After the cleanup, there are approximately 1.27 million busy trips.

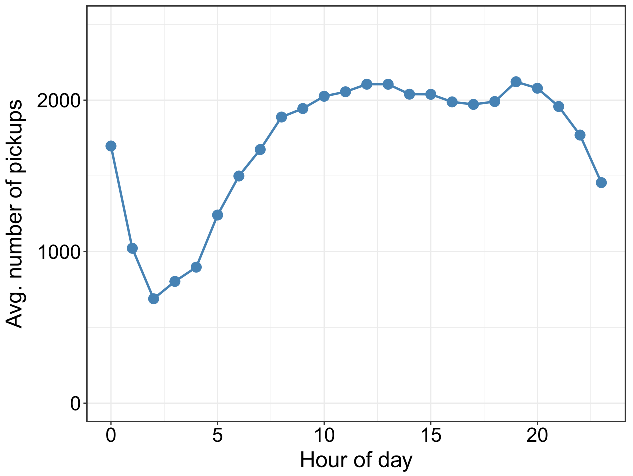

FIGURE 2. Average number of pickups in the study area considered in the Bangkok dataset in each hour of the day.

The average number of pickups in the study area in each hour of the day is illustrated in Fig. 2. It can be observed that the average number of pickups follows a typical pattern of hourly demand. Basically, the average number of pickups is small in the early morning hours (e.g., between 1 am and 4 am). It then increases during the daytime and drops during the evening hours.

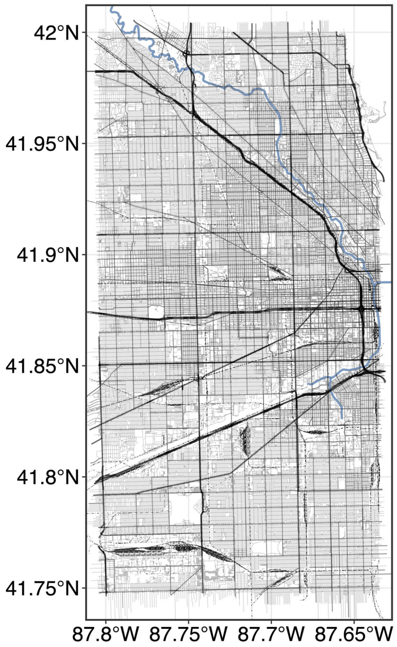

FIGURE 3. The study area considered in the Chicago dataset is shown inside the rectangular bounding box. (This map is created from the open data provided by OpenStreetMap [21] under the Open Database Licence [22].)

2) CHICAGO TAXI DATA

The second dataset considered in this study is publicly provided by the Chicago Data Portal [23]. Unlike the

TABLE 2. Four corners of the study area in the Chicago dataset.

Corner	Latitude	Longitude
Upper left	42.00016	-87.80400
Upper right	42.00016	-87.63488
Lower left	41.74788	-87.80400
Lower right	41.74788	-87.63488

Bangkok dataset, which contains only the raw GPS records of each taxi, the data in the Chicago dataset are already provided in a trip format. Important variables of each trip include the timestamp and the geolocations of the census tracts where the pickup and the drop-off occur. For privacy reasons, the actual pickup and drop-off locations are not revealed. Only the latitude and longitude of the census tract or the community area where pickup and drop-off occur are provided. The study area in the Chicago dataset is defined as the rectangular region shown in Fig. 3. The geolocations of the four corners of the study area are given in Table 2.

The total number of pickups in the study area in December 2019 is 315,464. The average number of pickups in the study area in each hour of the day is shown in Fig. 4. Similar to that observed in the Bangkok dataset, the average number of pickups follows a typical pattern of hourly demand. Basically, the average number of pickups is small in the early morning hours (e.g., between 12 am and 6 am). It then increases during the daytime and drops during the evening hours.

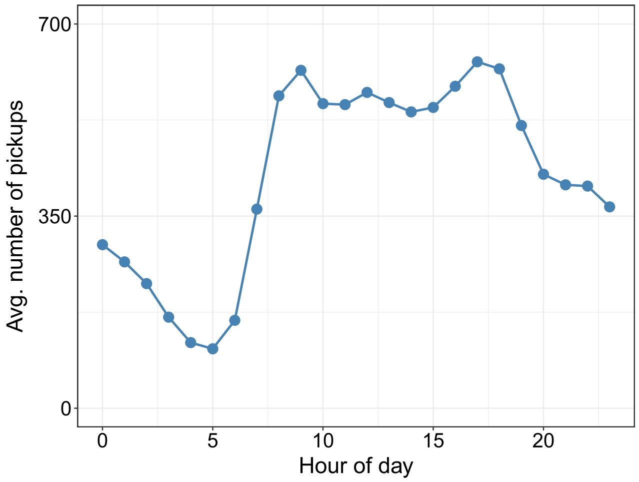

FIGURE 4. Average number of pickups in the study area considered in the Chicago dataset in each hour of the day.

B. EXPERIMENTAL PROCEDURES

In this section, we describe the approach used in analyzing the temporal distribution of the number of pickups. The main idea is to determine whether the number of pickups observed in a given space over time can be modeled by a known statistical distribution. A common technique used in identifying the temporal distribution is to count the number of pickups in a given region during an observation period and determine whether these values come from a particular distribution. In this study, we define a *test region* to be a square of size 1 km^2. The Bangkok study area can be divided into 3,417 test regions, and the Chicago study area can be divided into 392 test regions.

Due to the periodic nature of demand, the average number of pickups in each hour of the day varies, as shown in Fig. 2 and Fig. 4. The parameters of the distribution that describes the number of pickups in each hour will definitely be different. As a result, it is not logical to find a single statistical distribution to model the variation of pickups within a day. Instead, it is more rational to investigate if the number of pickups during the same hour of the day over different days follows any statistical distribution (e.g., poisson distribution). Thus, we observe how the number of pickups in a given hour of the day varies over different days for each test region. This is the same as the scenario where the poisson distribution is traditionally used to model customer arrivals at fixed points, such as taxi stands. The only difference is that we model the arrivals (pickups) in a fixed space (e.g., a square of size 1 km^2) instead of a fixed point. Let R_i be a test region i. Let T_j for $j \in \{0, 1, \ldots, 23\}$ be the hours of the day. For example, T_0 refers to the period between 12 am and 1 am, and T_{23} refers to the period between 11 pm and midnight. A pair of test region and the hour of the day (i.e., (R_i, T_j)) uniquely identifies a set of pickup samples in the region, which will be referred to as a *temporal test set*. Since there are 31 days in December, each temporal test set will consist of 31 pickup samples. There are 3,417 test regions in the Bangkok dataset and 24 hours in a day. As a result, there would be a total of 3,417 × 24 = 82,008 temporal test sets in the Bangkok dataset. Similarly, there are 392 test regions in the Chicago dataset. As a result, there would be a total of 392 × 24 = 9,408 temporal test sets in the Chicago dataset.

To characterize the distribution of the number of pickups in a temporal test set, we need to determine whether these samples are drawn from a particular hypothesized distribution. For example, we can ask if these 31 samples could have come from a poisson distribution. Since the number of pickups is a discrete value, a discrete probability distribution should be used as a hypothesized distribution. In addition, the distribution should be able to support the entire range of non-negative integers (i.e., $x \in \{0, 1, 2, \ldots\}$). The following three well-known discrete probability distributions fit this profile; therefore, they are used as the hypothesized distributions in this study.

1) *Poisson distribution:* Due to many of its properties that help simplify the analysis, a poisson distribution is often assumed in many studies, as discussed in the related work section. A discrete random variable X is a poisson random variable if its probability mass function (PMF) can be described as

$$P_X(x) = \begin{cases} \dfrac{\lambda^x}{x!} e^{-\lambda}, & x \in \{0, 1, 2, \ldots\} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where $\lambda > 0$ is the parameter of the distribution. A rather unique characteristic of the poisson distribution is that both its mean and its variance are equal to λ.

2) *Geometric distribution:* A geometric distribution typically models the number of Bernoulli trials before success. It can be regarded as a discrete counterpart of the exponential distribution. To our knowledge, it has not been used to model the temporal distribution of the number of pickups in any existing studies. A discrete random variable X is a geometric random variable if its PMF can be described as

$$P_X(x) = \begin{cases} (1-p)^x p, & x \in \{0, 1, 2, \dots\} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where p is the probability of success in each Bernoulli trial. Note that p is the only parameter of the geometric distribution.

3) *Negative binomial distribution:* A negative binomial distribution typically models the number of failures in a sequence of Bernoulli trials before a specified number of successes are observed. It was used to model the pickup count of the NYC taxis [7]. A discrete random variable X is a negative binomial random variable if its PMF can be described as

$$P_X(x) = \begin{cases} \binom{x+r-1}{x}(1-q)^x q^r, & x \in \{0, 1, 2, \dots\} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where r is the specified number of successes, and q is the probability of success in each Bernoulli trial. In contrast to the poisson distribution and the geometric distribution, the negative binomial distribution has two parameters (i.e., r and q).

In this study, we determine whether any of these three distributions can be used to model the number of pickups in each temporal test set.

A statistical procedure commonly used in determining whether a set of random samples comes from a hypothesized distribution is the goodness-of-fit test. Since the hypothesized distributions are discrete distributions, we resort to the chi-square goodness-of-fit test. The idea is to assume that the samples are drawn from a hypothesized distribution (e.g., poisson, geometric, and negative binomial) and see if the null hypothesis can be rejected at a desired statistical significance level. The null hypothesis and the alternative hypothesis in the chi-square goodness-of-fit test are given as follows.

H_0: samples are from the hypothesized distribution

H_1: samples are not from the hypothesized distribution

Essentially, the chi-square goodness-of-fit test determines if the difference between the values observed from the samples and the values expected from the hypothesized distribution is significant enough to reject the null hypothesis. Readers are referred to [24] for additional details on the chi-square goodness-of-fit test. In our experiment, we perform the chi-square goodness-of-fit test on each of the temporal test sets for each type of the hypothesized distributions. The parameters of the hypothesized distributions used in evaluating each temporal test set are estimated from the pickup samples in the set. More specifically, the parameter λ of the poisson distribution, the parameter p of the geometric distribution, and the pair of parameters (r, q) of the negative binomial distribution are estimated from the pickup samples in the temporal test set under investigation. The chi-square goodness-of-fit test results on these temporal test sets are presented and discussed in Section IV.

IV. RESULTS AND DISCUSSION

A. BANGKOK TAXI DATA

The chi-square goodness-of-fit test is performed on each temporal test set to determine whether the pickup samples in the set come from the hypothesized distributions. In this study, the null hypothesis is rejected at the 5% significance level. It is not meaningful to analyze the test sets that mostly have no pickups; therefore, only the temporal test sets where the mean number of pickups is at least 1 are considered. The temporal test sets that fail the chi-square test (i.e., the null hypothesis is rejected) and those that pass the chi-square test (i.e., the null hypothesis cannot be rejected) are identified.

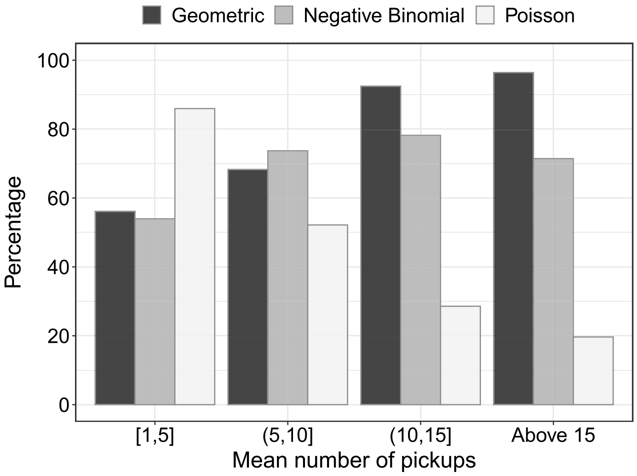

FIGURE 5. Percentage of the temporal test sets that pass the chi-square goodness-of-fit test for each type of hypothesized distributions. Four different ranges of the mean number of pickups are compared. The data are from the Bangkok taxi dataset.

The overall test result is shown in Fig. 5. The horizontal axis illustrates four ranges of the mean number of pickups in the temporal test sets. The first range represents the temporal test sets with the mean number of pickups between 1 and 5, while the last range represents the temporal test sets with the mean number of pickups above 15. The vertical axis illustrates the percentage of the temporal test sets in each range that pass the chi-square goodness-of-fit test. Basically, the height of each bar indicates the percentage of the temporal test sets that can be modeled by the hypothesized distribution. For example, for the temporal test sets with the mean number of pickups between 1 and 5, 86% of them can be modeled by the poisson distribution. In other words, 86% of the test sets in this range pass the chi-square test when the poisson distribution is used as the hypothesized distribution. In contrast, for the temporal test sets with the mean number of pickups above 15, only 20% of them can be modeled by the poisson distribution. Generally, it can be observed that when the mean number of pickups is extremely small, the poisson distribution can model the pickup samples really well. However, as the mean number of pickups increases, the poisson distribution becomes much less effective. On the contrary, as the mean number of pickups increases, the geometric distribution becomes increasingly more effective in modeling the variation of the pickup samples. In fact, more than 92% of the temporal test sets can be modeled by the geometric distribution when the mean number of pickups is between 10 and 15. Moreover, when the mean number of pickups is above 15, the geometric distribution can model more than 96% of the temporal test sets. Evidently, when the mean number of pickups is above 10, the geometric distribution is more effective than the poisson distribution and the negative binomial distribution.

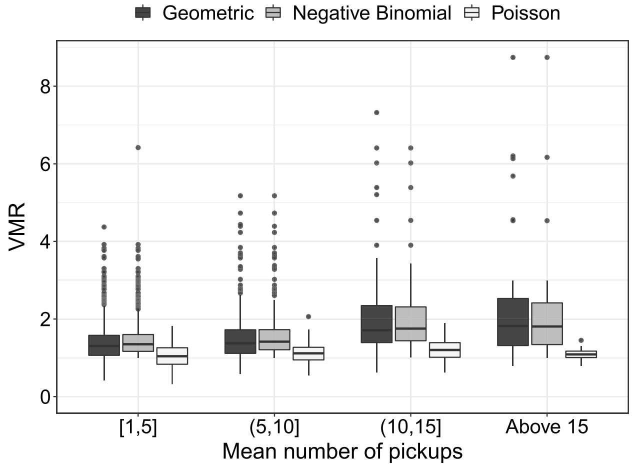

FIGURE 6. A box plot of variance-to-mean ratio of the temporal test sets that pass the chi-square goodness-of-fit test. The data are from the Bangkok taxi dataset.

It is essential to understand the characteristics of the pickup samples that each type of the hypothesized distributions can model. For instance, it is important to know why the poisson distribution can only model the pickup samples well when the mean number of pickups is extremely small. To characterize the pickup samples in each temporal test set, we use a quantitative variable called *variance-to-mean ratio* (VMR). It is basically a ratio between the variance and the mean of the pickup samples. A box plot of the VMR of the temporal test sets that pass the chi-square goodness-of-fit test is shown in Fig. 6. Clearly, the VMR values of the temporal test sets that the poisson distribution can model concentrate around 1. This is coherent with the fact that the VMR value of the poisson distribution is always equal to 1 (i.e., the variance and the mean are identical). In fact, the temporal test sets with the VMR values that are much larger than 1 cannot be modeled by the poisson distribution. On the other hand, the negative binomial distribution cannot model the temporal test sets with the VMR values that are smaller than 1. Finally, the geometric distribution covers the range of the VMR values that cannot be modeled by the poisson distribution and the negative binomial distribution. The VMR values in this range correspond to those that are much larger than 1 and those that are smaller than 1.

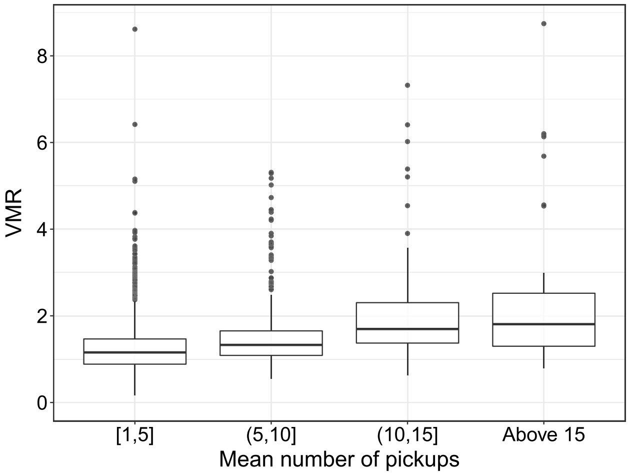

FIGURE 7. A box plot of variance-to-mean ratio of all the temporal test sets at different ranges of the mean number of pickups. The data are from the Bangkok taxi dataset.

To understand why the poisson distribution becomes less effective and the geometric distribution becomes more effective when the mean number of pickups increases, we investigate the VMR of all the temporal test sets at different values of the mean number of pickups. A box plot of the VMR of all the temporal test sets at different ranges of the mean number of pickups is shown in Fig. 7. Generally, it can be observed that the VMR values increase as the mean number of pickups increases. When the mean number of pickups is extremely small (i.e., between 1 and 5), most of the VMR values are close to 1. This is the main reason why the poisson distribution can model most of the temporal test sets in this range. However, as the mean number of pickups increases, the VMR values of the temporal test sets become much larger than 1. In this case, the poisson distribution is not effective in modeling the pickup samples. On the contrary, the geometric distribution, which can model the pickup samples with high VMR exceptionally well, becomes distinctly effective when the mean number of pickups increases.

A close look at the PMFs of these two distributions provides a theoretical justification for why the geometric distribution might be more suitable than the poisson distribution.

The PMF of the poisson distribution with parameter λ, where $\lambda > 0$, is as described in (1). Both its mean and variance are equal to λ. Therefore, the VMR value of the poisson distribution is always equal to $\lambda/\lambda = 1$. As a result, the poisson distribution will not be able to model the empirical samples with the VMR value that is much larger than 1, as observed in the empirical results. On the other hand, the PMF of the geometric distribution with parameter p, where $p \in [0, 1]$, is as described in (2). Its mean is equal to $(1 - p)/p$, and its variance is equal to $(1 - p)/p^2$. Therefore, the VMR value of the geometric distribution is always equal to $1/p$. As a result, the geometric distribution can model the empirical samples with much larger VMR values. Fig. 8 illustrates the VMR value of the geometric distribution as a function of the parameter p. Clearly, the geometric distribution can support an extensive range of VMR values. This theoretically explains why the geometric distribution is more suitable than the poisson distribution, especially when the empirical samples have high VMR values.

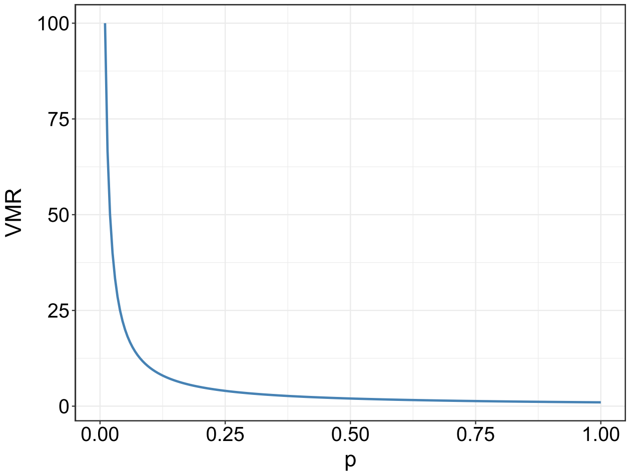

FIGURE 8. The variance-to-mean ratio of the geometric distribution as a function of the parameter p.

It is crucial to know if spatial resolution has any effect on the test results observed earlier. In other words, would changing the size of the test region alter the conclusions drawn from the result shown in Fig. 5? To investigate this issue, we repeat the experiment by changing the test region size from 1 km^2 to 4 km^2. The percentage of the temporal test sets that pass the chi-square goodness-of-fit test in the scenario where the test region size is 4 km^2 is shown in Fig. 9. Essentially, the same trend as noted in Fig. 5 is still observed here. The poisson distribution can only model the pickup samples well when the mean number of pickups is extremely small. It becomes much less effective when the mean number of pickups increases. On the contrary, the geometric distribution can model the pickup samples more effectively as the mean number of pickups increases. A similar result is also observed when the test region size is changed to 2 km^2. However, the result is omitted due to space limitations.

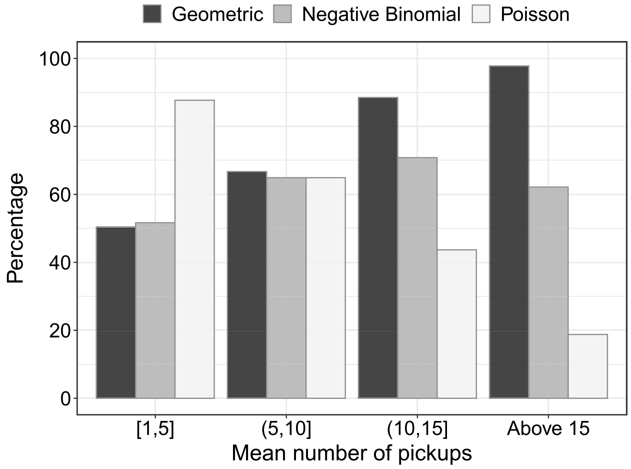

FIGURE 9. Percentage of the temporal test sets that pass the chi-square goodness-of-fit test for each type of hypothesized distributions in the scenario where the test region size is 4 km^2. The data are from the Bangkok taxi dataset.

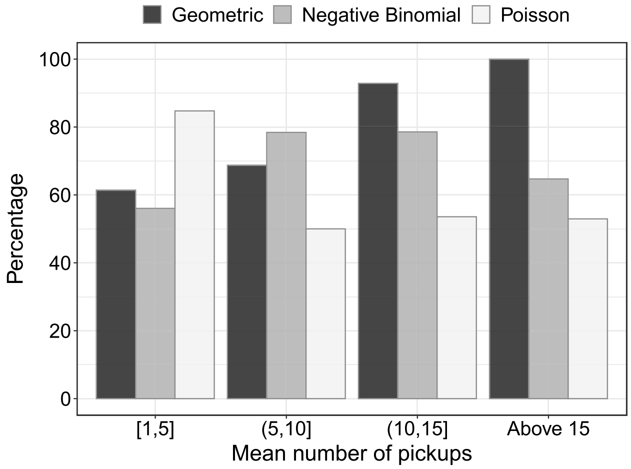

FIGURE 10. Percentage of the temporal test sets that pass the chi-square goodness-of-fit test for each type of hypothesized distributions in the scenario where the length of the observation period is 30 minutes. The data are from the Bangkok taxi dataset.

Next, we investigate if time resolution affects the test results observed earlier. In other words, would changing the length of the observation period alter the conclusions drawn from the result shown in Fig. 5? To examine this issue, we repeat the experiment by changing the length of the observation period from 1 hour to 30 minutes. The percentage of the temporal test sets that pass the chi-square goodness-of-fit test in the scenario where the length of the observation period is 30 minutes is shown in Fig. 10. In this scenario, we still see the same trend that we have observed in Fig. 5. As the mean number of pickups increases, the poisson distribution becomes less effective in modeling the pickup samples. On the contrary, as the mean number of pickups increases, the geometric distribution becomes increasingly more effective in modeling the pickup samples. A similar result is also observed when the length of the observation period is changed to 2 hours. However, the result is omitted due to space limitations.

In summary, we can conclude from the Bangkok taxi data that the poisson distribution can only model the pickup samples well when the mean number of pickups is extremely small. The main reason is that the pickup samples with small mean tend to have a VMR value close to 1, which fits the unique characteristic of the poisson distribution. However, as the mean number of pickups increases, the VMR value tends to be much higher than 1. In this case, the geometric distribution is more effective in modeling the pickup samples.

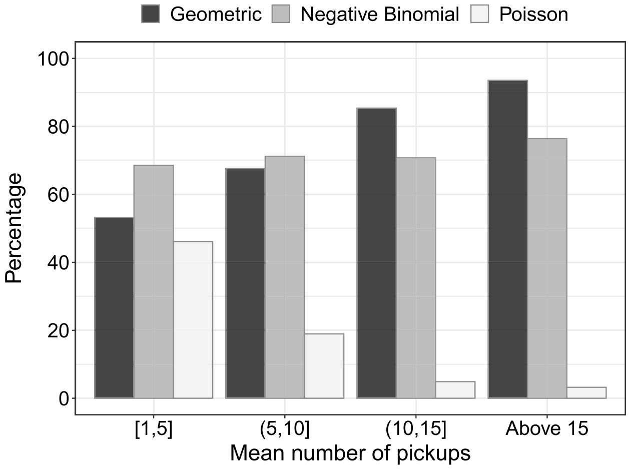

FIGURE 11. Percentage of the temporal test sets that pass the chi-square goodness-of-fit test for each type of hypothesized distributions. Four different ranges of the mean number of pickups are compared. The data are from the Chicago taxi dataset.

B. CHICAGO TAXI DATA

The same chi-square goodness-of-fit test procedure performed on the Bangkok dataset is also repeated on the Chicago dataset. The temporal test sets that fail the chi-square test (i.e., the null hypothesis is rejected) and those that pass the chi-square test (i.e., the null hypothesis cannot be rejected) are identified. The percentage of the temporal test sets that pass the chi-square test for each type of the hypothesized distributions is shown in Fig. 11. Similar to the Bangkok dataset, trend-wise, the effectiveness of the poisson distribution in modeling the pickup samples decreases as the mean number of pickups increases. Nonetheless, a subtle difference can be observed. Unlike the results observed in the Bangkok dataset, the percentage of the temporal test sets that can be modeled by the poisson distribution in this dataset is much smaller. This is due to the fact, which will be shown later, that most of the VMR values of the temporal test sets in this dataset are much larger than 1. This makes the poisson distribution unsuitable for modeling these pickup samples. Similar to the Bangkok dataset, however, we observe that the geometric distribution can model the pickup samples exceptionally well when the mean number of pickups increases. Indeed, when the mean number of pickups is greater than 10, the geometric distribution is more effective than the other two distributions.

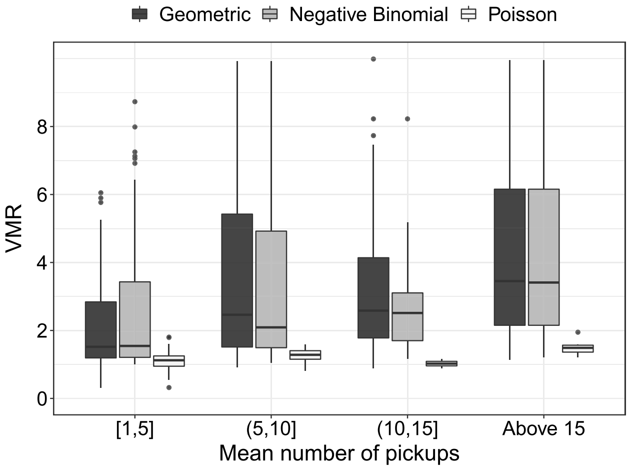

FIGURE 12. A box plot of variance-to-mean ratio of the temporal test sets that pass the chi-square goodness-of-fit test. The data are from the Chicago taxi dataset.

A box plot of the VMR of the temporal test sets that pass the chi-square goodness-of-fit test is shown in Fig. 12. Similar to the Bangkok dataset, it can be observed that the poisson distribution can only model the temporal test sets with the VMR values that concentrate around 1. The negative binomial distribution cannot model the test sets with the VMR values that are smaller than 1. Evidently, the geometric distribution can model the temporal test sets with a wider range of VMR values in comparison to the other two distributions.

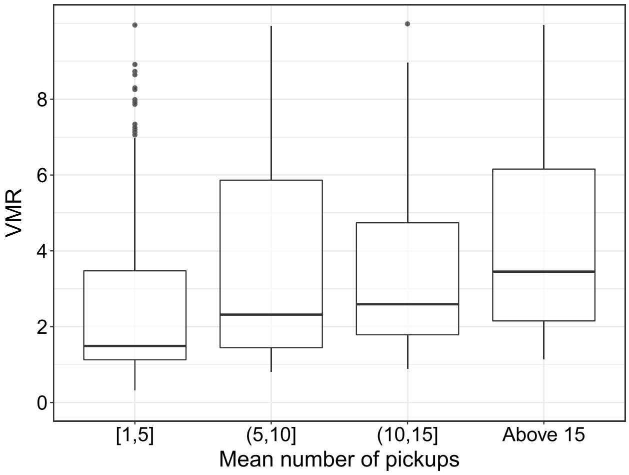

FIGURE 13. A box plot of variance-to-mean ratio of all the temporal test sets at different ranges of the mean number of pickups. The data are from the Chicago taxi dataset.

Analyzing the VMR values of the temporal test sets reveals why the poisson distribution is not effective in modeling the pickup samples across all ranges of the mean number of pickups in the Chicago dataset. A box plot of the VMR of all the temporal test sets at different ranges of the mean number of pickups is shown in Fig. 13. It can be observed that most of the VMR values of the temporal test sets, across all ranges of the mean number of pickups, are much larger than 1. This is the main reason why the poisson distribution is not able to model the pickup samples well across all ranges. Similar to the Bangkok dataset, trend-wise, we observe that the VMR values tend to increase as the mean number of pickups increases. This explains why the effectiveness of the geometric distribution increases while the effectiveness of the poisson distribution decreases as the mean number of pickups increases.

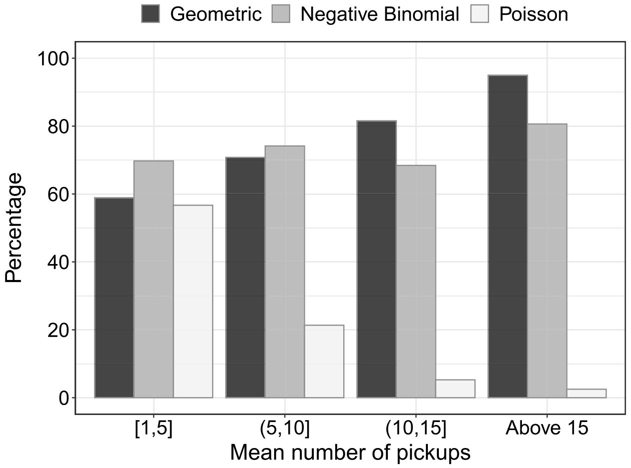

FIGURE 14. Percentage of the temporal test sets that pass the chi-square goodness-of-fit test for each type of hypothesized distributions in the scenario where the test region size is 4 km^2. The data are from the Chicago taxi dataset.

To investigate if spatial resolution has any effect on the conclusions drawn from the result shown in Fig. 11, we repeat the experiment by changing the test region size from 1 km^2 to 4 km^2. The percentage of the temporal test sets that pass the chi-square goodness-of-fit test in the scenario where the test region size is 4 km^2 is shown in Fig. 14. Basically, the same trend as noted in Fig. 11 is still observed here. The effectiveness of the poisson distribution decreases while the effectiveness of the geometric distribution increases as the mean number of pickups increases. A similar result is also observed when the test region size is changed to 2 km^2. However, the result is omitted due to space limitations.

Finally, to investigate if time resolution affects the conclusions drawn from the result shown in Fig. 11, we repeat the experiment by changing the length of the observation period from 1 hour to 30 minutes. The percentage of the temporal test sets that pass the chi-square goodness-of-fit test in the scenario where the length of the observation period is 30 minutes is shown in Fig. 15. Clearly, changing the time resolution does not alter the conclusions drawn from the result discussed earlier. In this scenario, we still observe the same trend that we have observed in Fig. 11. A similar result is also observed when the length of the observation period is changed to 2 hours. However, the result is omitted due to space limitations.

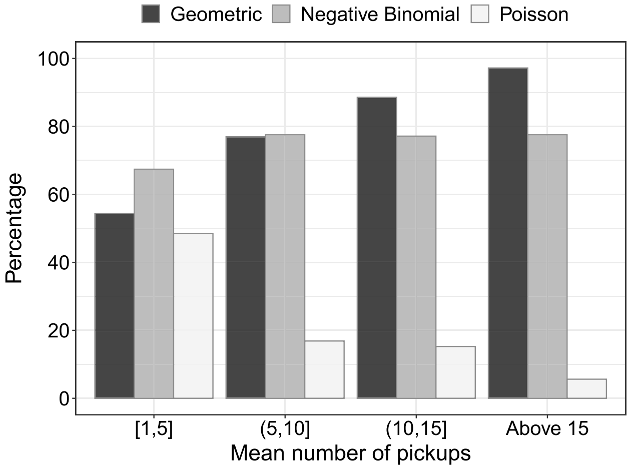

FIGURE 15. Percentage of the temporal test sets that pass the chi-square goodness-of-fit test for each type of hypothesized distributions in the scenario where the length of the observation period is 30 minutes. The data are from the Chicago taxi dataset.

In summary, the results observed in the Chicago dataset are similar to those observed in the Bangkok dataset. As the mean number of pickups increases, the poisson distribution becomes less effective. On the contrary, the geometric distribution becomes more effective as the mean number of pickups increases. In the scenario where the mean number of pickups is greater than 10, it is clear that the geometric distribution is more suitable for modeling the pickup samples than the other two distributions.

Taxi demand is subject to temporal effects (e.g., time of the day, weekdays, weekends) and spatial effects (e.g., commercial area, residential area, nightlife area). One certainly could analyze a conditional distribution of taxi demand given a combination of these conditions (e.g., a commercial area during the 9 am-10 am period on weekdays, a nightlife district during the 9 pm-10 pm period on weekends, etc.). However, this study aims to characterize the marginal probability distribution that represents taxi demand in general without conditioning on the type of days or land use.

In fact, this study provides a higher layer of abstraction beyond these specific conditions. We treat a temporal test set as a spatiotemporal unit where its characteristics are purely defined by the statistical traits of the demand in the unit, not by the type of days and land use. We demonstrate that the poisson distribution is appropriate for a test set with an extremely low mean number of pickups. On the contrary, the geometric distribution is more suitable for a test set with a higher mean number of pickups. Essentially, we use the statistical characteristics of the demand to discriminate the scenarios where each type of distribution would be applicable. Therefore, our results are also applicable to scenarios with specific conditions. For example, to select an appropriate probability distribution to model taxi demand in a commercial area from 9 am to 10 am on weekdays, one can first determine the mean number of pickups in the area from 9 am to 10 am on weekdays. If the mean number of pickups is high, the geometric distribution should be used. Similarly, to select an appropriate probability distribution to model taxi demand in a residential area from 10 pm to 11 pm on weekends, one can first determine the mean number of pickups in the area from 10 pm to 11 pm on weekends. If the mean number of pickups is extremely low, the poisson distribution could be used.

C. IMPLICATION OF THE RESULTS

Based on the empirical evidence shown in the previous sections, it is clear that the poisson distribution is only effective in modeling the temporal distribution of the number of pickups in the scenario where the mean number of pickups is extremely small. In most practical scenarios, the geometric distribution is more suitable. In this section, we will quantify the implication when the incorrect assumption (i.e., the poisson distribution) is used in lieu of the correct distribution (i.e., the geometric distribution). Essentially, we will evaluate the difference it makes when the incorrect assumption is presumed.

A typical metric commonly used for measuring the difference between two probability distributions is the Kullback-Leibler (KL) divergence. The KL divergence between two discrete probability distributions, $P_X(x)$ and $Q_X(x)$, is defined as [25].

$$D(P\|Q) = -\sum_x P_X(x)\log\left(\frac{Q_X(x)}{P_X(x)}\right). \tag{4}$$

The KL divergence basically measures the expected log difference between the two distributions. Particularly in our case, we will use the KL divergence to quantify the difference between the poisson distribution and the geometric distribution.

Let $P_X(x)$ be the PMF of the poisson distribution with mean λ, and let $Q_X(x)$ be the PMF of the geometric distribution with the same mean. Then, $P_X(x)$ is as given in (1), and $Q_X(x)$ is as given in (2) with parameter $p = 1/(\lambda+1)$. In other words, $Q_X(x)$ is

$$Q_X(x) = \begin{cases} \left(1-\frac{1}{\lambda+1}\right)^x\left(\frac{1}{\lambda+1}\right), & x \in \{0,1,2,\ldots\} \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Substituting $P_X(x)$ and $Q_X(x)$ into (4), the KL divergence can be expressed as

$$D(P\|Q) = (\lambda+1)\log(\lambda+1) - \lambda - \sum_{x=0}^{\infty}\frac{\lambda^x}{x!}e^{-\lambda}\log x!\,. \tag{6}$$

Unfortunately, the exact expression for the KL divergence given in (6) does not have a closed-form solution, and thus it has to be obtained numerically. Nonetheless, an approximated solution can be derived. This can be done by rewriting the KL divergence in terms of the difference between the cross entropy of $P_X(x)$ and $Q_X(x)$ and the entropy of $P_X(x)$, which is [25]

$$D(P\|Q) = H(P,Q) - H(P) \tag{7}$$

where

$$H(P,Q) = -\sum_x P_X(x)\log Q_X(x) \tag{8}$$

is the cross entropy of $P_X(x)$ and $Q_X(x)$, and

$$H(P) = -\sum_x P_X(x)\log P_X(x) \tag{9}$$

is the entropy of $P_X(x)$.

The cross entropy of $P_X(x)$ and $Q_X(x)$ has an exact closed-form solution, which can be written as

$$H(P,Q) = (\lambda+1)\log(\lambda+1) - \lambda\log\lambda. \tag{10}$$

The entropy of $P_X(x)$ is the entropy of the poisson distribution, which unfortunately does not have a closed-form solution. However, it can be approximated as [26]

$$H(P) \approx \frac{1}{2}\log(2\pi e\lambda) - \frac{1}{12\lambda} + O\left(\frac{1}{\lambda^2}\right). \tag{11}$$

Finally, substituting (10) and (11) into (7), the KL divergence between the poisson distribution and the geometric distribution can be approximated as

$$D(P\|Q) \approx (\lambda+1)\log(\lambda+1) - \lambda\log\lambda - \frac{1}{2}\log(2\pi e\lambda) + \frac{1}{12\lambda}. \tag{12}$$

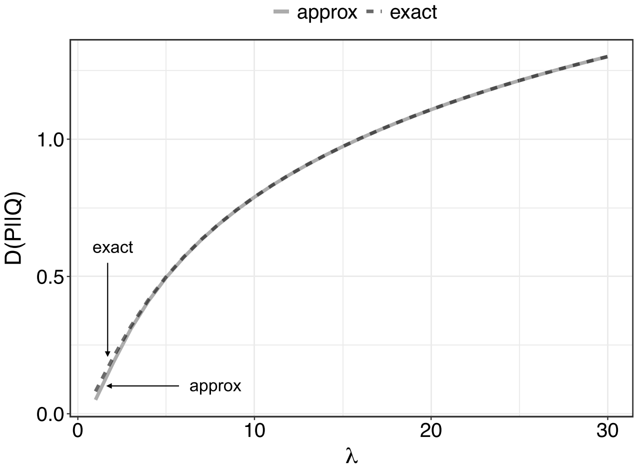

FIGURE 16. The KL divergence between the poisson distribution and the geometric distribution as a function of the mean of the distributions (i.e., λ).

In Fig. 16, the KL divergence between the poisson distribution and the geometric distribution is shown as a function of the mean of the distributions (i.e., λ). The values obtained numerically from the exact solution in (6) are shown with the dashed line, while those obtained from the approximation in (12) are shown with the solid line. First, it can be observed that the values of $D(P\|Q)$ obtained from the approximation and those obtained from the exact solution are remarkably close. This verifies that the approximation is valid. Second, it can be observed that the KL divergence between the two distributions increases as the mean of the distributions, λ, increases. Even at a small value of λ, the two probability distributions already exhibit a significant divergence. For instance, at $\lambda = 10$, the KL divergence between the two distributions is around 0.79, which means that they differ by a factor of $e^{0.79} \approx 2$. This basically implies that, at $\lambda = 10$, the probability of the number of pickups estimated by the poisson distribution will differ from the actual probability by a factor of 2, on average. The probability of the number of pickups estimated by the poisson distribution deviates from that of the actual distribution (i.e., the geometric distribution) by a larger factor as the mean number of pickups increases.

D. LIMITATIONS

In this section, we discuss a few limitations of this work.

1) OBSERVED PICKUPS AND ACTUAL DEMAND

Like most related work on taxi demand, this study uses the observed number of pickups as a proxy for actual demand. In practice, the actual demand is likely larger than the observed number of pickups. In fact, the observed number of pickups only represents a scaled or subsampling version of the actual demand. However, the main objective of this study is not to obtain the numerical value of actual demand. Instead, we focus on characterizing the statistical distributions that can be used to model the actual demand. The type of distribution is determined from the scaled version of the actual demand. The actual demand follows the same type of distribution as its scaled version but with different distribution parameters (e.g., with a larger mean). Indeed, proper distribution parameters for the actual demand distribution or an appropriate scaling factor must be obtained from a field experiment. Nonetheless, the pickup data suffice for characterizing the type of demand distribution.

2) REPRESENTATIVENESS OF DATA

The taxi data investigated in this study are from Bangkok, Thailand, and Chicago, USA. These data represent taxi operations in a typical urban environment in Southeast Asian and North American cities. The two datasets are independent, and both confirm our hypothesis on the temporal distribution of taxi demand. Nonetheless, the distribution of taxi demand in other regions could differ from what we have investigated here. It remains open for future study.

3) SPECIAL EVENTS

This study only concentrates on characterizing the statistical distribution of taxi demand in a normal situation. Special events, such as concerts, conferences, and holidays, may affect demand distribution. For example, these events can create a surge in demand in a particular location. Characterizing the distribution of demand under these exceptional circumstances could be interesting, but it is beyond the scope of this study.

V. CONCLUSION

In this study, we characterize the temporal distribution of the number of pickups based on the real taxi trip data in Bangkok and Chicago. A chi-square goodness-of-fit test is performed on three types of hypothesized distributions, which are the poisson distribution, the negative binomial distribution, and the geometric distribution. It is shown that the poisson distribution is only effective in modeling the temporal distribution of the number of pickups in the scenario where the mean number of pickups is extremely small (i.e., between 1 and 5 pickups). Its effectiveness decreases immensely when the mean number of pickups increases. On the contrary, the effectiveness of the geometric distribution increases as the mean number of pickups increases. In fact, the geometric distribution can model the temporal distribution of the number of pickups exceptionally well in most scenarios (i.e., when the mean number of pickups is greater than 10).

To select an appropriate distribution for a particular area and observation period in practice, one must first determine the mean number of pickups in that area and period. This can be achieved by observing the historical data. For example, to determine a suitable distribution for modeling the number of pickups between 8 am and 9 am in a desired area, one can first observe the historical pickup data and determine the mean number of pickups in such area and period. If the mean number of pickups is extremely small, use the poisson distribution. Otherwise, use the geometric distribution.

The poisson arrival assumption has been used extensively in most analyses on mobility demand, especially in demand prediction. However, we have shown empirically in this study that this assumption is not likely valid. In fact, most of the time, this fundamental assumption does not hold. A more appropriate distribution to use, on the contrary, is the geometric distribution. It would be interesting to see how the existing demand analyses would change under the assumption that the temporal distribution of the number of pickups is a geometric distribution. This is a subject worth investigating in future research. Nonetheless, at the fundamental level, we have shown in this study that the probability of the number of pickups estimated by the poisson distribution is significantly different from that of the geometric distribution. The divergence between the two distributions gets more prominent as the mean number of pickups increases.

Lastly, the results presented in this study open up a few new research directions. First, new theoretical demand prediction models based on a geometric probability distribution can be investigated. Second, new queuing models and simulation models can be developed based on the new discovery that the arrivals of pickup requests follow a geometric distribution. Finally, new strategies for rebalancing demand and supply in an on-demand mobility service system can also be explored.

REFERENCES

[1] L. Moreira-Matias, J. Gama, M. Ferreira, J. Mendes-Moreira, and L. Damas, "Predicting taxi–passenger demand using streaming data," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 3, pp. 1393–1402, Sep. 2013.

[2] 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.

[3] M. Hyland, F. Dandl, K. Bogenberger, and H. Mahmassani, "Integrating demand forecasts into the operational strategies of shared automated vehicle mobility services: Spatial resolution impacts," *Transp. Lett.*, vol. 12, no. 10, pp. 671–676, Nov. 2020.

[4] B. Jäger, M. Wittmann, and M. Lienkamp, "Analyzing and modeling a city's spatiotemporal taxi supply and demand: A case study for Munich," *J. Traffic Logistics Eng.*, vol. 4, no. 2, pp. 147–153, Dec. 2016.

[5] N. J. Yuan, Y. Zheng, L. Zhang, and X. Xie, "T-finder: A recommender system for finding passengers and vacant taxis," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 10, pp. 2390–2403, Oct. 2013.

[6] A. K. Menon and Y. Lee, "Predicting short-term public transport demand via inhomogeneous Poisson processes," in *Proc. ACM Conf. Inf. Knowl. Manag.*, Nov. 2017, pp. 2207–2210.

[7] C. Yang and E. J. Gonzales, "Modeling taxi demand and supply in New York city using large-scale taxi GPS data," in *Seeing Cities Through Big Data*, P. Thakuriah, N. Tilahun, and M. Zellner, Eds. Berlin, Germany: Springer, 2017, pp. 405–425.

[8] J. Butkevič ius and A. Juozapavicius, "The methodology of modelling taxi rank service," *TRANSPORT*, vol. 18, no. 4, pp. 153–156, Jun. 2003.

[9] R. Zhang and R. Ghanem, "Demand, supply, and performance of street-hail taxi," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 10, pp. 4123–4132, Oct. 2020.

[10] W. Zhang, H. Honnappa, and S. V. Ukkusuri, "Modeling urban taxi services with E-hailings: A queueing network approach," *Transp. Res. Proc.*, vol. 38, pp. 751–771, Jan. 2019.

[11] X. Zheng, X. Liang, and K. Xu, "Where to wait for a taxi?" in *Proc. ACM SIGKDD Int. Workshop Urban Comput.*, Aug. 2012, pp. 149–156.

[12] G. Qi, G. Pan, S. Li, Z. Wu, D. Zhang, L. Sun, and L. T. Yang, "How long a passenger waits for a vacant taxi - large-scale taxi trace mining for smart cities," in *Proc. IEEE Int. Conf. Green Comput. Commun. IEEE Internet Things IEEE Cyber, Phys. Social Comput.*, Aug. 2013, pp. 1029–1036.

[13] D. K. George and C. H. Xia, "Fleet-sizing and service availability for a vehicle rental system via closed queueing networks," *Eur. J. Oper. Res.*, vol. 211, no. 1, pp. 198–207, May 2011.

[14] S. Rajendran and J. Shulman, "Study of emerging air taxi network operation using discrete-event systems simulation approach," *J. Air Transp. Manag.*, vol. 87, Aug. 2020, Art. no. 101857.

[15] S. Ghosh, P. Varakantham, Y. Adulyasak, and P. Jaillet, "Dynamic repositioning to reduce lost demand in bike sharing systems," *J. Artif. Intell. Res.*, vol. 58, pp. 387–430, Feb. 2017.

[16] A. Wallar, M. van der Zee, J. Alonso-Mora, and D. Rus, "Vehicle rebalancing for mobility-on-demand systems with ride-sharing," in *Proc. IEEE Int. Conf. Intell. Robots and Syst. (IROS)*, Oct. 2018, pp. 4539–4546.

[17] R. Zhang and M. Pavone, "Control of robotic mobility-on-demand systems: A queueing-theoretical perspective," *Int. J. Robot. Res.*, vol. 35, nos. 1–3, pp. 186–203, Jan. 2016.

[18] T. Babicheva, M. Cebecauer, D. Barth, W. Burghout, and L. Kloul, "Empty vehicle redistribution with time windows in autonomous taxi systems," *ACM/IMS Trans. Data Sci.*, vol. 2, no. 1, pp. 1–22, Jan. 2021.

[19] M. Grzegorczyk and M. Shafiee Kamalabad, "Comparative evaluation of various frequentist and Bayesian non-homogeneous Poisson counting models," *Comput. Statist.*, vol. 32, no. 1, pp. 1–33, Mar. 2017.

[20] Intelligent Traffic Information Center (iTIC) Foundation. (Sep. 2021). *Historical Raw Vehicles and Mobile Probes Data in Thailand*. iTIC Open Data Archives. [Online]. Available: https://org.iticfoundation.org/download

[21] *OpenStreetMap*. Accessed: Mar. 7, 2023. [Online]. Available: https://www.openstreetmap.org

[22] *Open Data Commons*. Accessed: Mar. 7, 2023. [Online]. Available: https://opendatacommons.org

[23] City of Chicago. Feb. 2023. *Taxi Trips 2019*. Chicago Data Portal. [Online]. Available: https://data.cityofchicago.org/Transportation/Taxi-Trips-2019/h4cq-z3dy

[24] D. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, 4th ed. Boca Raton, FL, USA: Chapman & Hall, 2007.

[25] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.

[26] R. J. Evans and J. Boersma, "The entropy of a Poisson distribution: Problem 87–6," *SIAM Rev.*, vol. 30, no. 2, pp. 314–317, 1988.

SOOKSAN PANICHPAPIBOON (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, USA, in 2000, 2002, and 2006, respectively. He is currently a Full Professor with the School of Information Technology, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand. His research interests include intelligent transportation systems, vehicular networks, mobile sensors, and performance modeling. He received the Asia-Europe Meeting DUO-Thailand Fellowship, in 2007, and the Doctoral Dissertation Award from the National Research Council of Thailand, in 2011.

KAVEPOL KHUNSRI is currently pursuing the integrated bachelor's and master's degree in information technology with the School of Information Technology, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand. His research interests include 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 Technology (ECTI-CON 2021).