



Article

# Modeling the Distribution of Human Mobility Metrics with Online Car-Hailing Data—An Empirical Study in Xi'an, China

Chaoyang Shi <sup>1</sup> , Qingquan Li <sup>2</sup>, Shiwei Lu <sup>3,4,\*</sup> and Xiping Yang <sup>5</sup> 

<sup>1</sup> School of Civil and Hydraulic Engineering, Huazhong University of Science and Technology, Wuhan 430074, China; cyshi@hust.edu.cn

<sup>2</sup> Shenzhen Key Laboratory of Spatial Smart Sensing and Services, Shenzhen University, Shenzhen 518060, China; liqq@szu.edu.cn

<sup>3</sup> School of Architecture and Urban Planning, Huazhong University of Science and Technology, Wuhan 430074, China

<sup>4</sup> Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources, Shenzhen 518034, China

<sup>5</sup> School of Geography and Tourism, Shaanxi Normal University, Xi'an 710119, China; xpyang@snnu.edu.cn

\* Correspondence: lusw@hust.edu.cn

**Abstract:** Modeling the distribution of daily and hourly human mobility metrics is beneficial for studying underlying human travel patterns. In previous studies, some probability distribution functions were employed in order to establish a base for human mobility research. However, the selection of the most suitable distribution is still a challenging task. In this paper, we focus on modeling the distributions of travel distance, travel time, and travel speed. The daily and hourly trip data are fitted with several candidate distributions, and the best one is selected based on the Bayesian information criterion. A case study with online car-hailing data in Xi'an, China, is presented to demonstrate and evaluate the model fit. The results indicate that travel distance and travel time of daily and hourly human mobility tend to follow Gamma distribution, and travel speed can be approximated by Burr distribution. These results can contribute to a better understanding of online car-hailing travel patterns and establish a base for human mobility research.

**Keywords:** mobility metrics; distribution fitting; Gamma distribution; Burr distribution; online car-hailing



**Citation:** Shi, C.; Li, Q.; Lu, S.; Yang, X. Modeling the Distribution of Human Mobility Metrics with Online Car-Hailing Data—An Empirical Study in Xi'an, China. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 268. <https://doi.org/10.3390/ijgi10040268>

Academic Editor: Wolfgang Kainz

Received: 27 February 2021

Accepted: 14 April 2021

Published: 17 April 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

The modeling of human mobility is an emergent research area. Studying the regularity and characteristics of human spatiotemporal mobility is of great significance in many fields, such as urban planning [1,2], traffic forecasting [3], and epidemic prevention [4,5].

When modeling human mobility, it is common to consider the probability distribution function (PDF) of its metrics (e.g., travel distance, travel time, and travel speed). It is generally accepted that daily and hourly human mobility metrics have a representative distribution [6]. Modeling the distributions of these metrics is fundamental, necessary, and beneficial for studying underlying travel patterns and establishing a base for human mobility research.

Recently, with the rapid development of information and communication technique (ICT) and location-based service (LBS) applications, online car-hailing equipped with Global Positioning System (GPS) plays an increasingly important role in people's daily travel activities. As an important data source, online car-hailing platforms (e.g., Uber, Lyft, and Didi Chuxing) generate a large amount of accurate location data. Unlike traditional survey data, cell phone datasets, wireless network traces, and taxi locations data, online car-hailing data are characterized by high-quality, high-resolution and a large-scale, which reflect the detailed spatiotemporal trajectory and actual origin and destination of people's

travels. Therefore, online car-hailing has established a rich and solid data foundation for distribution modeling, thus bringing new opportunities and challenges to further understand people's travel behavior and intra-urban mobility.

To our knowledge, previous studies have proposed quite a few patterns of human mobility, such as levy flight model [7,8], power law distribution [9–13], exponential distribution [14–18], lognormal distribution [19–22], Weibull distribution [23], Pareto distribution [24], and Gamma distribution [22]. Brockmann et al. observed that human travel distance exhibits a power law distribution by analyzing the statistical properties of bank notes' circulation, and human travel trajectories may be approximated as Lévy flights (heavy tailed random walk) [7]. This observation was confirmed by Rhee et al. using Global Positioning System (GPS) traces collected from volunteers, showing a non-negligible probability of high displacement trips and a long pause-time between trips [8]. Despite the randomness indicated by Lévy flight models, a power law with an exponential cutoff can be used to approximate the displacement distribution of human trajectories obtained from mobile phone datasets [9,11], GPS traces [12] and online location-based social networks [13]. However, Kang et al., Jiang et al., and Liang et al. pointed out that an exponential distribution can be used to approximate taxis' travel displacement and travel time, instead of power law [14–18]. Furthermore, by analyzing a taxi-trace dataset, Wang et al. found that displacement tends to follow exponential distribution, and travel time is approximated by lognormal distribution [19].

Although the findings mentioned above provide a beneficial reference on mining human mobility, they mostly focus on a single model, which may not fit all data well. Zheng et al. found that a fusion function, based on exponential power law and a truncated Pareto distribution, represents travel time distribution best [24]. Bazzani et al. studied the GPS data of private cars in Florence, Italy, and found that the single-trip length follows an exponential behavior in short distance scale but favors a power law distribution for trips longer than 30km [18]. Csáji et al. and Zhang et al. found that exponential distribution is not appropriate for travel distances, and log-normal distribution provides reasonable fits [20,21]. Plötz et al. used Weibull, Gamma, and lognormal distributions to fit individual daily driving distances, and found that Weibull and lognormal most often perform better than Gamma, and the Weibull distribution fits most data but not all [23]. Kou and Cai analyzed the distributions of travel distance and travel time, and found that both of them follow a lognormal distribution in larger bike sharing systems, while the distribution for smaller systems varies among Weibull, Gamma, and lognormal [22].

To sum up, according to various datasets, many empirical studies have demonstrated that mobility metrics may be fitted with several meaningful distributions, such as Lévy flight models, exponential, power-law, lognormal, Gamma, Weibull, and Rayleigh. However, based on a real large-scale car-hailing trajectory dataset, can a single or mixed model achieve a good fit for all data? It remains to be further explored. In addition, the above studies focused on modeling the distribution of the human mobility with simple or overall data, while ignoring the variability of PDF along with day of week and time of day. Is the distribution type of mobility metrics in different time granularity different from that of overall data? If yes, how will it vary, and can it be described by a general distribution? This has aroused the interest of many scholars. Therefore, our research, based on a real large-scale car-hailing trajectory dataset, is indispensable, and gains valuable insight into human mobility patterns.

To fill this gap, this paper aims to model the distribution of human mobility metrics in different time granularity. Specifically, three metrics (travel distance, travel time, and travel speed) are introduced to explore massive trajectory data collected in Xi'an, China. For each mobility metric, several candidate distributions are compared based on model selection criteria, and the best one is selected. The statistical distributions of daily trip data are analyzed first, and they show the characteristic of skewed distribution. More granularly, hourly distributions are further evaluated, and a general distribution for each mobility metric is determined.

The remainder of this paper is organized as follows. Section 2 briefly introduces the online car-hailing dataset and carries out the basic analysis. Section 3 describes the trip metrics, the fitting distributions, and the method of model selection. Section 4 presents the result and analysis. Section 5 discusses the findings. Finally, Section 6 provides conclusions and recommendations for further research.

## 2. Data Collection and Basic Analysis

### 2.1. Data Description

The adopted trajectory data were generated by about 18,000 online car-hailing trips in Xi'an, China, from 1 October 2016 to 30 November 2016. Vehicle trajectories were composed of high-resolution GPS points, which were recorded every 2–4 s. Accordingly, an online car-hailing trajectory is a sequence of GPS sampling points with five fields. The vehicle ID and order ID were desensitized to protect privacy. The “Timestamp” indicated when the data would be recorded, which was the UTC time. “Latitude” and “Longitude” provided location information of online car-hailing.

Let  $Tr_i^j = (p_1^{ij}, p_2^{ij}, \dots, p_N^{ij})$  denote the trajectory of the  $j$ th trip of vehicle  $i$ , where  $p_n^{ij} = (x, y, t)_n^{ij}$  is the  $n$ th point of the sequence ( $n = 1, 2, \dots, N$ ).  $(x, y)_n^{ij}$  denotes the location and  $t_n^{ij}$  the timestamp, respectively. Given a trajectory,  $t_1^{ij} < t_2^{ij} < \dots < t_N^{ij}$ . For a vehicle, the origin and destination (OD) locations are the first and last sampling points of a trip. It makes sense to define  $p_O^{ij} = p_1^{ij}$  and  $p_D^{ij} = p_N^{ij}$ . Hence, each OD trip can be simplified to be a vector from  $p_O^{ij}$  to  $p_D^{ij}$ .

A road network consists of a set of nodes, directed links, and allowed movements. Each node is a geographical location representing a network intersection, which can be either signalized or non-signalized. A link is defined to be the road section from its tail node to head node. The relative position denotes the ratio of a sampling point relative to the link start node, which ranges  $[0, 1]$ . For example, the value 0, 0.5, and 1 of the relative position represent the beginning, middle and end of a link.

### 2.2. Data Preprocessing

To model the travel distance distribution, the map matching (MM) and the path inference algorithm proposed by Chen et al. were first used [25]. As shown in Table 1, the original latitude and longitude were converted into geodetic coordinates, which could be used directly to calculate the travel displacement. Secondly, the relative position of the sampling point on the link was also calculated, and the UTC time was converted to the time of day (0–86400 s). Thirdly, the pick-up and drop-off points were extracted to calculate the trip metrics (e.g., travel time, travel displacement, travel distance).

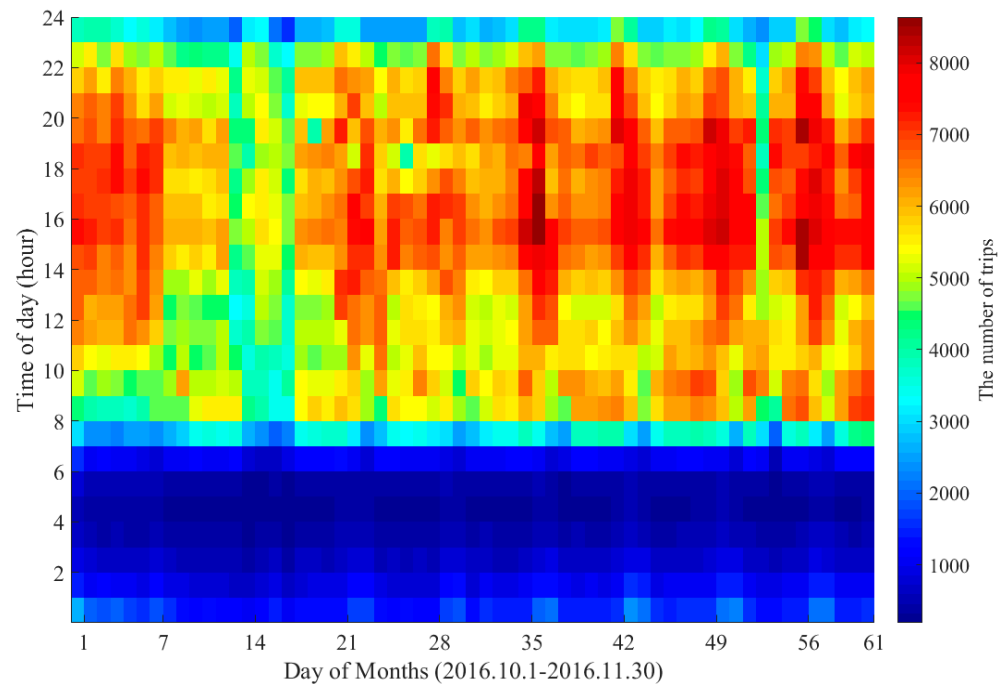
**Table 1.** Data section after processing.

Vehicle ID	Order ID	Timestamp	X-Coordinate	Y-Coordinate	Link ID	Relative Position
1000001	20000001	25,414	491,832.85	3,787,627.44	42915	0.3116
1000001	20000001	25,417	491,848.40	3,787,654.13	42915	0.8538
1000001	20000001	25,420	491,876.34	3,787,700.57	20558	0.5930

Finally, data cleaning is an essential task, because some of the trip records were not suitable for use in this study. Considering travel costs, few passengers travel by online car-hailing when travel time and distance are very short or long [26,27]. In addition, travel speed should be within a reasonable range. Therefore, the following conditions resulted in exclusion of trip records from the study data: (1) travel distance and displacement between origin and destination of less than 300m; (2) travel time less than 1min or longer than 2 h; (3) average travel speed below 5 km/h or in excess of 80km/h [28].

In terms of the total trips of two months, 6,203,848 trips were obtained from 6,584,397 original trips after data cleaning, which means that about 6% of the trips were filtered

out. From the perspective of daily trips, the average order availability was 94.22%, which fluctuated between 93.21% and 94.85%. More commonly, the study period was discretized into 1464 (24\*61) 1 h intervals for further analysis of residents' hourly trips. The hourly trip quantity ranged from 192 to 8636, as shown in Figure 1. On the whole, the number of trips during the day was much higher than at night, which is in line with human mobility. After all, human mobility during the day is more active, important, and meaningful. In addition, the number of hourly trips between 00:00 and 07:00 may be less than 2000, but it was sufficient for distribution fitting.



**Figure 1.** Distribution of hourly trip quantity.

### 3. Trip Metrics and Data Fitting

#### 3.1. Trip Metrics

In the existing studies, due to the lack of map matching, travel distance is usually replaced by displacement or Manhattan distance of the origin and destination. However, travel distance here refers to the length of the actual path traveled by the OD trip in road-networks. A path is composed of a series of successive links, and its length is the sum of length of links included in the OD trip. It is worth noting that the links where the origin and destination (OD) are located may not be traveled through. Based on the map matching and path inference results, the travel distance  $d_i^j$  (TD) is calculated as:

$$d_i^j = (1 - r_O^{i,j}) \cdot d_O^{i,j} + \sum_{k=2}^{M-1} d_k^{i,j} + r_D^{i,j} \cdot d_D^{i,j} \quad (1)$$

where  $M$  is the number of links included in the trip.  $r_O^{i,j}$  and  $r_D^{i,j}$  denote the ratio of trip OD relative to the link start node, which ranges  $[0, 1]$ .  $d_O^{i,j}$  and  $d_D^{i,j}$  are the link length where the trip OD are located.

Travel time is another important metric and is closely tied to travel distance. Travel time means time elapsed from the origin to destination, and is influenced by real-time traffic conditions, weather conditions, driver's driving habits, etc. As an important indicator for

analyzing human mobility, travel time reflects accessibility and traffic conditions. For a trajectory of the trip  $j$  of vehicle  $i$ , travel time  $t_i^j$  (TT) is defined as:

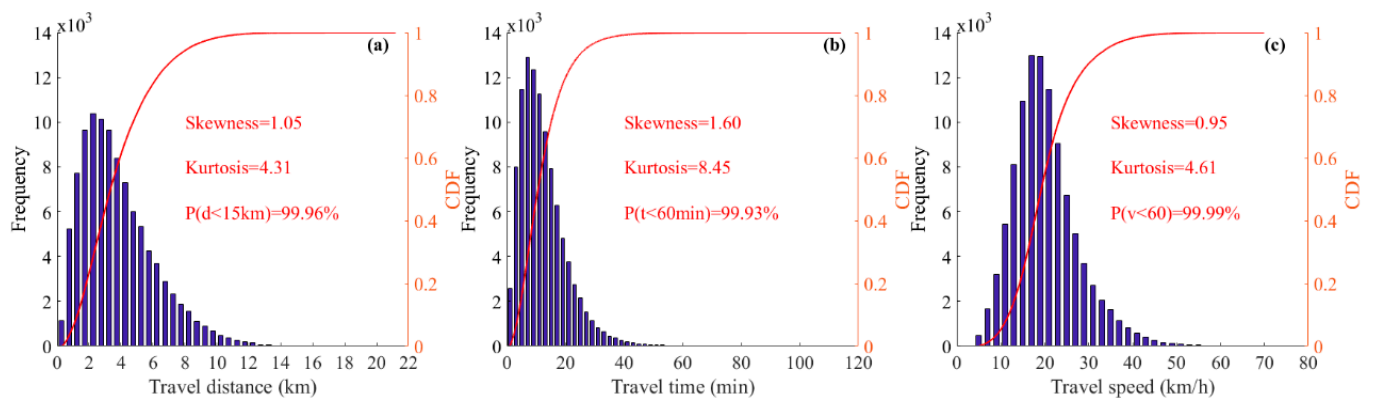
$$t_i^j = t_N^{ij} - t_1^{ij} = t_D^{ij} - t_O^{ij} \quad (2)$$

To understand the relation between the metrics described above, travel speed is another important feature. The average travel speed  $v_i^j$  (TS) is defined as:

$$v_i^j = d_i^j / t_i^j \quad (3)$$

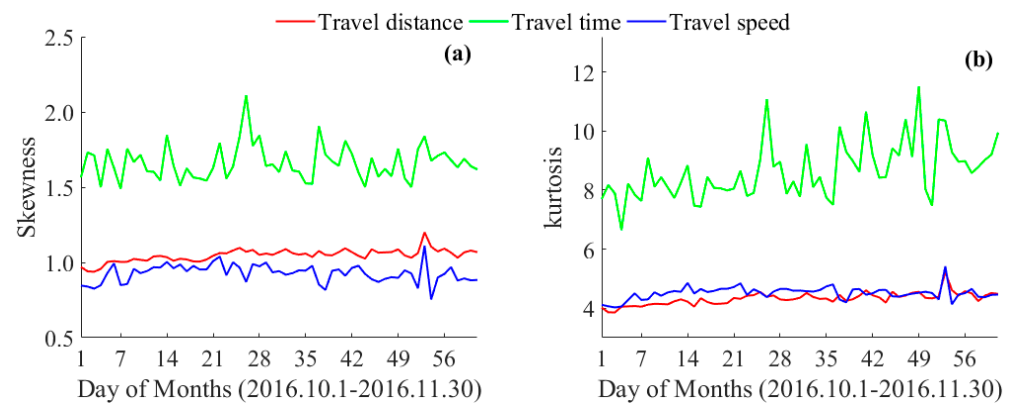
### 3.2. The Fitting Distribution

The fitting function selection is to identify the most appropriate distribution which is supported by the actual trip data. In the existing literature, exponential, (truncated) power-law, lognormal, Gamma, Weibull, and Burr distributions are commonly used distributions for fitting the above trip metrics [6,11,15,19,21,29–32]. However, not all of the above distributions apply to the data in this study. In order to narrow the range of candidate distributions, one day of the two-month data was randomly selected to analyze the characteristics of trip metrics. Figure 2 shows the frequency distribution histograms and the cumulative distribution functions (CDF) of travel distance, travel time, and travel speed, respectively. Based on the shape of these histograms, it can be found that these data show a significant right skew, which is also proved by the skewness calculation (i.e., 1.05, 1.60, 0.95). In addition, the CDF is very close to one before reaching the maximum. For example, the probability of travel distance within 15km is as high as 99.96%, and 99.93% trips have a travel time less than 1h. The excessive kurtosis indicates that the data are too concentrated, possibly due to the existence of extreme values.



**Figure 2.** Distribution of trip metrics on Thursday, November 3, 2016: (a) distribution of travel distance; (b) distribution of travel time; and (c) distribution of travel speed.

To further understand the shape of the data in detail, the skewness and kurtosis of the daily data are shown in Figure 3. The daily skewness is greater than 0.6, ranging from 0.75 to 2.1. This indicates that all data show a right skew, especially travel time data. Meanwhile, kurtosis greater than three indicates the steepness of the distribution, which proves the possibility of narrowing the spread range of trip metrics. More commonly, hourly skewness and kurtosis are, respectively, 1.07 and 5.83, which also indicate that the hourly trip data are highly likely to conform to the skewed distribution.



**Figure 3.** Daily variation of skewness and kurtosis for 61 days: (a) skewness; and (b) kurtosis

Based on the above analysis, five common skewed distributions—lognormal, Gamma, Weibull, Burr, and Rayleigh—are selected as candidate distributions to fit the daily and hourly data. The probability density functions (PDF) of these distributions are defined in the following formulas. Assuming that the variable  $x$  follows a lognormal distribution, its PDF can be expressed as

$$f(x) = \frac{1}{x \cdot \sigma \sqrt{2\pi}} \cdot e^{\left(\frac{-(\ln x - \mu)^2}{2\sigma^2}\right)} \quad (4)$$

where  $\mu$  and  $\sigma$  denote the mean and standard deviation of the natural logarithm of the variable  $x$ . The mathematical expectation and the variance are respectively  $E(x) = e^{\mu + \sigma^2/2}$  and  $Var(x) = (e^{\sigma^2} - 1) \cdot e^{2\mu + \sigma^2}$ .

The Gamma distribution with shape and scale parameters  $\alpha$  and  $\beta$  is

$$f(x) = \frac{1}{\beta^\alpha \cdot \Gamma(\alpha)} \cdot x^{\alpha-1} \cdot e^{-\frac{x}{\beta}} \quad (5)$$

where  $\Gamma(\cdot)$  presents the Gamma function, and  $\Gamma(1) = 1$ . The mean and variance are calculated with  $E(x) = \alpha \cdot \beta$ ,  $Var(x) = \alpha \cdot \beta^2$ . It should be noted here that the exponential and  $\chi^2$  distributions are special cases of the Gamma distribution. For example, when the shape parameter  $\alpha$  is 1, the Gamma distribution is an exponential distribution with the parameter  $1/\beta$ .

The Weibull distribution is defined as

$$f(x) = \frac{k}{\lambda} \cdot \left(\frac{x}{\lambda}\right)^{k-1} \cdot e^{-\left(\frac{x}{\lambda}\right)^k} \quad (6)$$

where  $k > 0, \lambda > 0$  are the shape and scale parameters, respectively. The mean and variance are  $E(x) = \lambda \cdot \Gamma\left(1 + \frac{1}{k}\right)$  and  $Var(x) = \lambda^2 \cdot \left[\Gamma\left(1 + \frac{2}{k}\right) - \Gamma\left(1 + \frac{1}{k}\right)^2\right]$ . The PDF of 3-parameter version of the Burr distribution is

$$f(x) = \frac{c \cdot k}{\alpha} \cdot \left(\frac{x}{\alpha}\right)^{c-1} \cdot \left(1 + \left(\frac{x}{\alpha}\right)^c\right)^{-(k+1)} \quad (7)$$

where  $\alpha > 0$  is a scaling parameter,  $c > 0$  and  $k > 0$  are shape parameters. The mean and variance are calculated by  $E(x) = \frac{\alpha \cdot k \cdot \Gamma\left(k - \frac{1}{c}\right) \cdot \Gamma\left(1 + \frac{1}{c}\right)}{\Gamma(1+k)}$  and  $Var(x) = \frac{\alpha^2 \cdot k \cdot \Gamma\left(k - \frac{2}{c}\right) \cdot \Gamma\left(1 + \frac{2}{c}\right)}{\Gamma(1+k)} - (E(x))^2$ .

Rayleigh's PDF has merely one parameter to estimate, making it popular for representing the distribution of trip metrics due to its simplicity. Its distribution is a special



case of the Weibull distribution in which the value of the shape parameter is 2. The PDF of Rayleigh distribution with a parameter  $\alpha$  is

$$f(x) = \frac{x}{\alpha^2} \cdot e^{-\frac{x^2}{2\alpha^2}} \quad (8)$$

The mean and variance are  $E(x) = \sqrt{\frac{\pi}{2}} \cdot \alpha$  and  $Var(x) = \frac{4-\pi}{2} \cdot \alpha^2$ , respectively. In the literature, the parameters are optimized by the maximum likelihood estimation (MLE), and detailed inference can refer to Clauset et al. [33].

### 3.3. Model Selection

In order to evaluate the fitted model between the actual data and candidate distribution, two fundamentally different methods, null hypothesis tests and model selection criteria, are frequently used to select appropriate model or the best model [34]. Theoretically, they can achieve both model fit and complexity. The Kolmogorov–Smirnov (K–S) test is commonly employed to evaluate the goodness of fit of the candidate distributions [6,22,29,35]. However, the K–S test is suitable for small samples. When the data are too large, the critical value for rejection is very small, and the result often rejects the null hypothesis. The K–S test may reject all the candidate distributions; however, it may also consider multiple distributions as acceptable. Considering the large quantity of daily and hourly trips shown in Figure 1, the K–S test is not suitable for this situation.

In addition, the Akaike information criterion (AIC) can provide another decision-making method [15,16,19]. The AIC score is a function of its maximized log-likelihood ( $L_i$ ) and the number of estimated parameters ( $K_i$ ) for each candidate model  $i$ , and is calculated by

$$AIC_i = -2 \cdot \ln L_i + 2 \cdot K_i \quad (9)$$

Generally, the model with the smallest AIC is preferred. In this study, small sample unbiased AIC is not considered due to the large number of daily or hourly trips. The number of hourly trips changes from a few hundred to nearly ten thousand, so it is important to find the distribution applicable to different periods. Therefore, the Bayesian information criterion (BIC, also referred to as Schwarz criterion SC) is further used for model selection. The BIC is structurally similar to the AIC, but includes a penalty term on sample size ( $N$ ), and tends to favor simpler models, particularly as the sample size increases.

$$BIC_i = -2 \cdot \ln L_i + K_i \cdot \ln N \quad (10)$$

The BIC is on a relative scale. The BIC difference  $\Delta_i = BIC_i - BIC_{\min}$  ( $BIC_{\min} = \min_{i \in \{1,2,\dots,n\}} \{BIC_i\}$ ) allows for an immediate ranking of the  $n$  candidate models [36]. The larger the BIC difference for a model, the less probable that it is the best model. More specifically, the Akaike weight  $w_i$  [37] represents the normalization of the relative likelihood (i.e.,  $e^{-\Delta_i/2}$ ) of the models.

$$w_i = \frac{e^{-\Delta_i/2}}{\sum_{j=1}^n e^{-\Delta_j/2}} \quad (11)$$

The Akaike weights are very useful for assessing model selection uncertainty. The model with the largest Akaike weight should be selected as the best distribution.

### 3.4. Evaluation of Distribution Fitness

In order to show how close the theoretical and empirical frequency distributions are, accuracy is evaluated using the following statistical indicators: coefficient of determination ( $R^2$ ), mean absolute error (MAE), mean absolute percentage error (MAPE), probability outside the predicted interval (POPI), probability outside the observed interval (POOI), and POPI-POOI-based criterion (PPC).

For many transportation applications, it is meaningful to construct an interval at a given confidence level from the fitting distribution. The accuracy of confidence interval represents the integrated accuracy of both the predicted mean and STD. In addition, percentiles are also an effective method of evaluating the accuracy of mean and STD. In this study, mean absolute error (MAE) and mean absolute percentage error (MAPE) of percentiles were extended as follows,

$$MAPE = \frac{100\%}{n} \cdot \sum_{i=1}^n \frac{|pt_i^{obs} - pt_i^{fit}|}{pt_i^{obs}} \quad (12)$$

$$MAE = \frac{1}{n} \cdot \sum_{i=1}^n |pt_i^{obs} - pt_i^{fit}| \quad (13)$$

where  $n$  is the number of percentiles, and  $pt_i^{obs}$ ,  $pt_i^{fit}$  are the observed percentiles obtained from the field survey and fitting distribution. Smaller  $MAPE$  and  $MAE$  indicate higher accuracy of the fitting distribution.

Two metrics were adopted to evaluate the accuracy of estimated or predicted distribution: probability outside the predicted interval (POPI) and probability outside the observed interval (POOI) [38,39]. The POPI measures the percentage of observed data outside the predicted confidence interval, while the POOI measures the percentage of predicted distribution outside the observed confidence interval. Let  $l_{fit} = \Phi_{fit}^{-1}(\alpha/2)$  and  $u_{fit} = \Phi_{fit}^{-1}(1 - \alpha/2)$  be the lower and upper bounds of the fitting distribution, respectively, at confidence level  $1 - \alpha$ , where  $\Phi_{fit}^{-1}(\cdot)$  is the inverse cumulative distribution function (CDF) of the fitting distribution. Mathematically, POPI is defined as

$$POPI = (1 - \sum_{i=1}^N c_i/N) \times 100\% \quad (14)$$

where  $c_i = 1$  if sample  $\in [l_{fit}, u_{fit}]$ , otherwise  $c_i = 0$ . The  $POPI$  value ranges from 0 to 1. The smaller  $POPI$  indicates capture of larger proportion of observed data, i.e., higher accuracy of the fitting distribution. As noted by Shi et al. [38,39], this  $POPI$  metric is very useful, but tends to exhibit bias for situations of wide fitting intervals due to large STD errors.

As an alternative, POOI metric is the percentage of predicted distribution outside the observed travel time interval. Let  $l_{obs}$  and  $u_{obs}$  denote the lower and upper bounds of the observed interval, respectively, at confidence level  $1 - \alpha$ . Then

$$POOI = (1 - (\Phi_{fit}(u_{obs}) - \Phi_{fit}(l_{obs}))) \times 100\% \quad (15)$$

$\Phi_{fit}(\cdot)$  denotes the CDF of the fitting distribution. POOI also ranges [0, 1], and larger POOI indicates lower fitting interval accuracy, because a larger proportion is outside the observed interval. Therefore, the POPI and POOI matrices are complementary for evaluating the accuracy of the fitting distribution. The closer the two to  $\alpha$ , the better the model fit. The bigger the POPI, the smaller the corresponding POOI. Therefore, a measure is required to balance the deviation of POPI and POOI from  $\alpha$ . With regard to this argument, the following POPI-POOI-based criterion (PPC) is proposed for the comprehensive evaluation of POPI and POOI:

$$PPC = \frac{1}{2} \cdot (|POPI - \alpha| + |POOI - \alpha|) \times 100\% \quad (16)$$

Generally, the smallness of PPC is an indication of the goodness of the constructed confidence interval (simultaneously achieving high model fit).

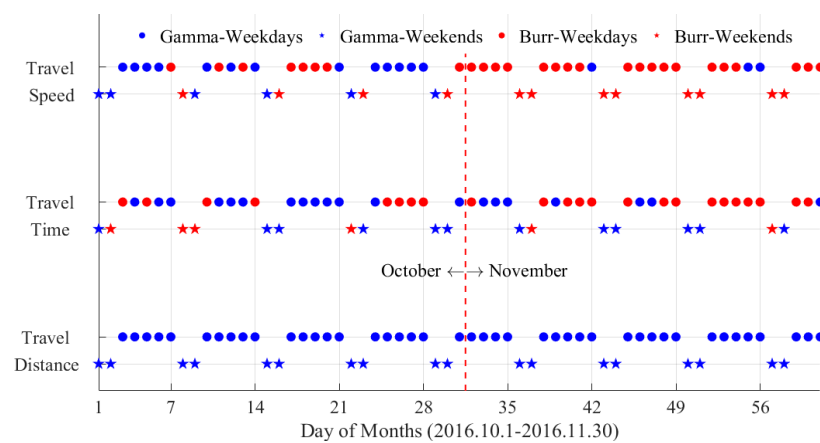


#### 4. Results of the Best-Fit Distribution

This section reports the fitting results of the trip metrics using the candidate skewed distributions. The best-fit distributions of daily trip metrics are first shown. Then, we further analyze the best-fit results of hourly trip metrics, such as travel distance, travel time, and travel speed. Finally, we attempt to identify a general distribution for each trip metric, and to estimate its parameters.

##### 4.1. Best-Fit Distribution of Daily Trip Metrics

Figure 4 shows the best-fit distributions of daily trip metrics for 61 days. Overall, only two of the candidate distributions, Gamma and Burr distributions (represented by blue and red), are suitable for fitting these trip metrics. Gamma distribution performs best for travel distance, and can uniformly fit all daily data, as shown in blue in Figure 4. However, Gamma distribution does not fit all travel time or speed data well. In total, 47.54% (29 out of 61) of travel time data and 63.93% (39 out of 61) of travel speed data follow the Burr distribution (depicted in red).

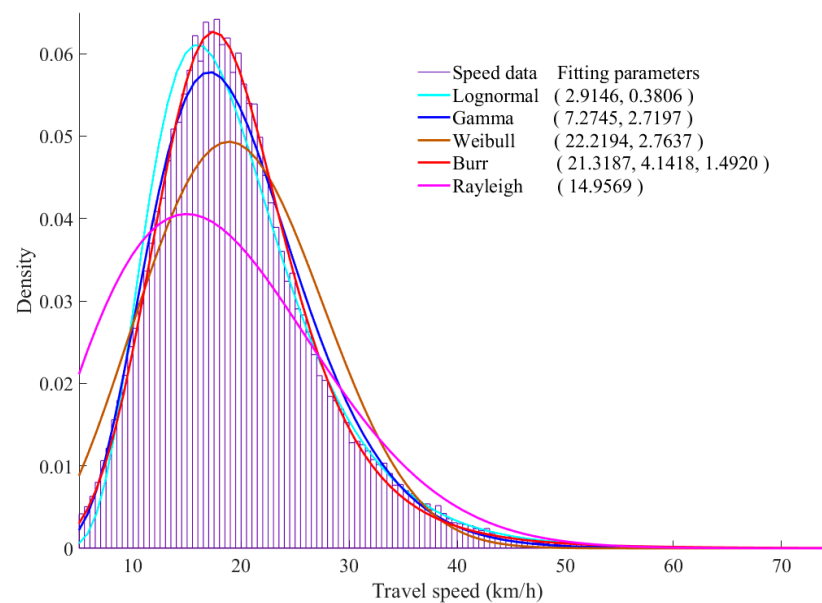


**Figure 4.** Best-fit distribution of daily trip metrics for 61 days.

The fitting results for weekdays and weekends in Figure 4 are distinguished by two different markers (dot and star). For travel distance, data on weekday and weekend can be well fitted with the same distribution. However, for travel time data at weekends, a third of them are subject to Gamma distribution, the rest to Burr distribution, and the weekend's speed data are the opposite. Meanwhile, only 37.21% (16 out of 43) of travel speed data on weekdays obey Gamma distribution, which further decreases to 13.64% (3 out of 22) on November weekdays.

The above analysis also shows that, due to uniform and simple fitting distribution, travel distance is more straightforward for analyzing residents' mobility patterns. In comparison, travel time and speed are relatively complex metrics due to uncertain distribution types.

In addition, the mean BIC weights of travel distance, travel time, and travel speed are 1, 0.9996, and 0.9960, respectively, which indicate low uncertainty of fitting results. The smallest BIC weight occurs in the travel speed data on October 13, and the fitting distributions and the detailed parameters are shown in Figure 5. It can be found from the figure that the Burr distribution is the most consistent with the speed data, followed by the Gamma distribution. It is noteworthy that the Burr distribution is more complex than the Gamma distribution. The likelihood of achieving a better fit of the more complex model is significantly greater than that of the simpler model, but the model fit and complexity should be considered comprehensively.



**Figure 5.** Fitted distributions for the observed travel speed data on October 13.

Table 2 shows more parameters in the model selection. The observed mean and standard variation (STD) are very similar to several estimates. The commonly used evaluation indices, such as the mean absolute percent error (MAPE) and the root mean square error (RMSE), are also difficult when it comes to identifying the dominant distribution. Furthermore, the log-likelihood of Burr distribution is slightly bigger than that of Gamma distribution, which quantitatively proves that the Burr distribution has a better model fit. When the BIC takes model complexity into account, the gap narrows to 2 ( $\Delta = 499,118 - 499,116 = 2$ ), which means that the benefit of improved model fit outweighs the cost of added model complexity. This tiny advantage is clearly distinguished in the BIC weight, which makes the model selection more explicit. It can be determined that the best-fit model is Burr distribution.

**Table 2.** Model selection for travel speed data on October 13 ( $\mu = 19.78, \sigma = 7.48$  km/h).

Distribution	Log-Likelihood ( $L_i$ )	BIC	The BIC weight	Mean	STD
Lognormal	−249,789	499,601	0	19.83	7.83
Gamma	−249,548	499,118	0.2412	19.78	7.34
Weibull	−253,174	506,369	0	19.78	7.74
Burr	−249,541	499,116	0.7588	19.81	7.73
Rayleigh	−259,304	518,620	0	18.75	9.80

#### 4.2. Best-Fit Distribution of Hourly Travel Distance

In order to further investigate the hourly distributions of trip metrics, the study period is discretized into 1464 (24\*61) 1 h intervals. Figure 6 shows the best-fit distributions of hourly travel distance. The best-fit distributions are distinguished by five different markers, of which the Gamma distribution accounts for 93.10%—far higher than the other four distributions. Between 07:00 and 24:00, the proportion of Gamma distribution rises to 99.32%, which further demonstrates the advantage of Gamma distribution in fitting travel distance data. During the times 00:00–07:00, the optimal distribution varies with hours and days, and is chaotically represented in five different markers. Only 77.99% of the data still obey Gamma distribution, while 14.52% for Weibull distribution, 6.79% for Rayleigh distribution, and less than 1% for lognormal distribution. This may be due to the small sample size at night.

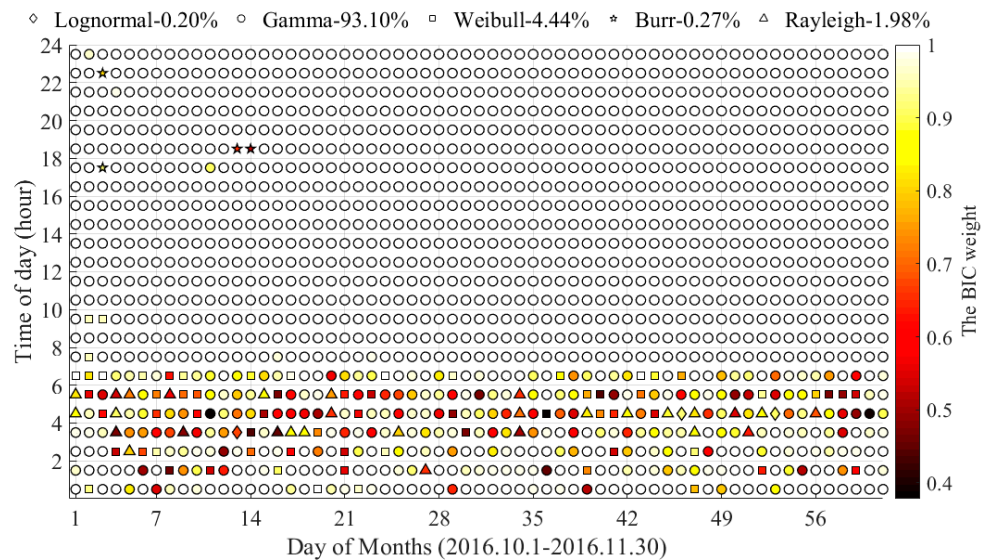


Figure 6. The fitting distributions of hourly travel distance.

In addition, the BIC weights represented in different colors range from 0.38 to 1. The darker the color, the smaller the BIC weight. It can be seen from Table 3 that most of the BIC weights between 07:00 and 24:00 are very close to 1. Their mean BIC weight is 0.9986, indicating the high reliability in the model selection. However, the mean BIC weight from 00:00 to 7:00 is 0.8540, which indicates that there is some uncertainty in the fitting results. More specifically, the mean BIC weights of the lognormal, Weibull, and Rayleigh distributions are 0.8362, 0.7094, and 0.7210, respectively. The high uncertainty may be caused by small sample size, because fewer people travel at night. Meanwhile, lower weights also mean that the best and suboptimal fitting distributions may both be applicable to the data. In conclusion, Gamma distribution may also be applicable to all the travel distance data.

Table 3. Statistics of best-fit distributions for travel distance.

Period	Index	Mean	Lognormal	Gamma	Weibull	Burr	Rayleigh
00:00–24:00	Percentage		0.20%	93.10%	4.44%	0.27%	1.98%
	BIC weight	0.9564	0.8362	0.9736	0.7207	0.7281	0.7210
00:00–07:00	Percentage		0.70%	77.99%	14.52%	0	6.79%
	BIC weight	0.8540	0.8362	0.8927	0.7094	0	0.7210
07:00–24:00	Percentage		0	99.32%	0.29%	0.39%	0
	BIC weight	0.9986	0	0.9998	0.9539	0.7281	0

### 4.3. Best-Fit Distribution of Hourly Travel Time

Figure 7 shows the best-fit distributions of hourly travel time, which are distinguished by different markers. As shown by the black circles and red squares, Gamma and Burr distributions are the dominant distribution types, with proportions of 76.23% and 20.56%, respectively. In Table 4, larger BIC weights (0.9430 and 0.8982) of these two distributions show a significant advantage over other three distributions when fitting 96.79% of the data. Meanwhile, the fitting results have higher reliability.

On the other hand, the other three distributions, shown by blue dots, green triangles, and red stars in Figure 7, account for less than 4%, and appear mainly at night (02:00–7:00). At the same time, their average BIC weights are only 0.7515, 0.6296, and 0.7853, respectively. These lower BIC weights indicate the higher uncertainty of the Lognormal, Weibull, and Rayleigh distributions in fitting the data, which are likely to be replaced by Gamma or Burr distribution.

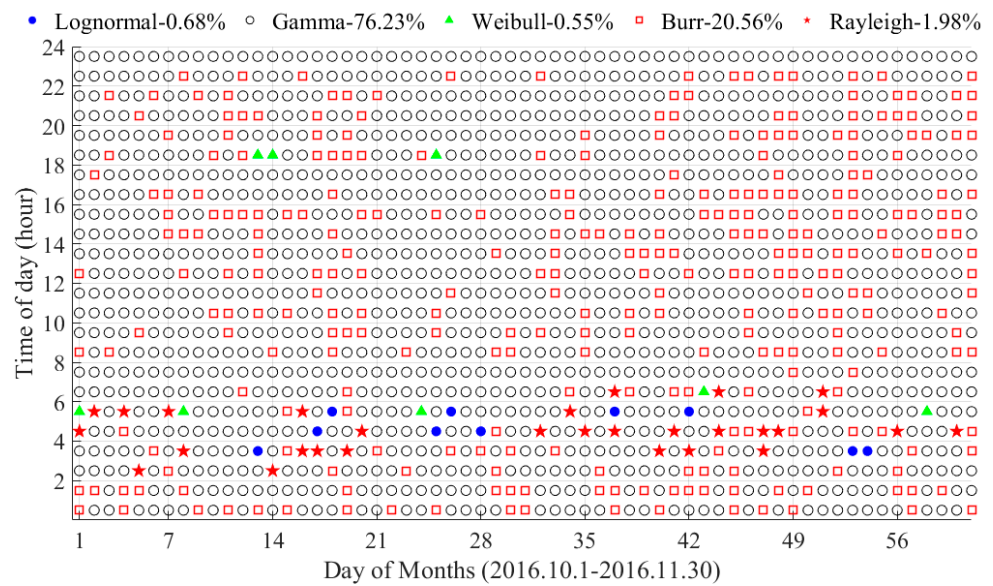


Figure 7. The fitting distributions of hourly travel time.

Table 4. Statistics of best-fit distributions for travel time.

Period	Index	Mean	Lognormal	Gamma	Weibull	Burr	Rayleigh
00:00-24:00	Percentage		0.68%	76.23%	0.55%	20.56%	1.98%
	BIC weight	0.9283	0.7515	0.9430	0.7446	0.8982	0.7853
02:00-07:00	Percentage		2.34%	71.43%	1.17%	18.27%	6.79%
	BIC weight	0.8525	0.7515	0.8800	0.6296	0.8669	0.7853
National Day holiday	Percentage		0	82.74%	0.60%	13.69%	2.98%
	BIC weight	0.9459	0	0.9541	0.5342	0.9352	0.7854

During the National Day holiday, 82.74% of travel time data follow Gamma distribution with a mean BIC weight of 0.9541. However, only 13.69% obey the Burr distribution, with an average BIC weight of 0.9352. This means that travel time data for the National Day holiday are more inclined to the Gamma distribution than the Burr distribution.

#### 4.4. Best-Fit Distribution of Hourly Travel Speed

Figure 8 shows the fitting results of hourly travel speed. The best-fit distributions are distinguished by rhombus, circle, square, and star. On the whole, Burr distribution accounts for 85.11% (1246 out of 1464) of the best-fit distributions, with a mean BIC weight of 0.9830, showing an absolute advantage and high reliability. Similarly, the lognormal and Gamma distributions have higher BIC weights but lower ratios. In contrast, the fitting results for only 6 1 h intervals (0.41%) are consistent with Weibull distribution, with an average BIC weight of 0.7007. This means that Weibull distribution is not suitable for speed data.

In addition, some clustering features can be found in non-dominant best-fit distributions. For example, 77.94% (53 out of 68) of the lognormal distributions occurs during the daily evening peak, with a mean BIC weight of 0.9986. About 40% of the Gamma distributions exists during the National Day holiday, with a mean BIC of 0.9246. In conclusion, Burr distribution is dominant in fitting travel speed data.

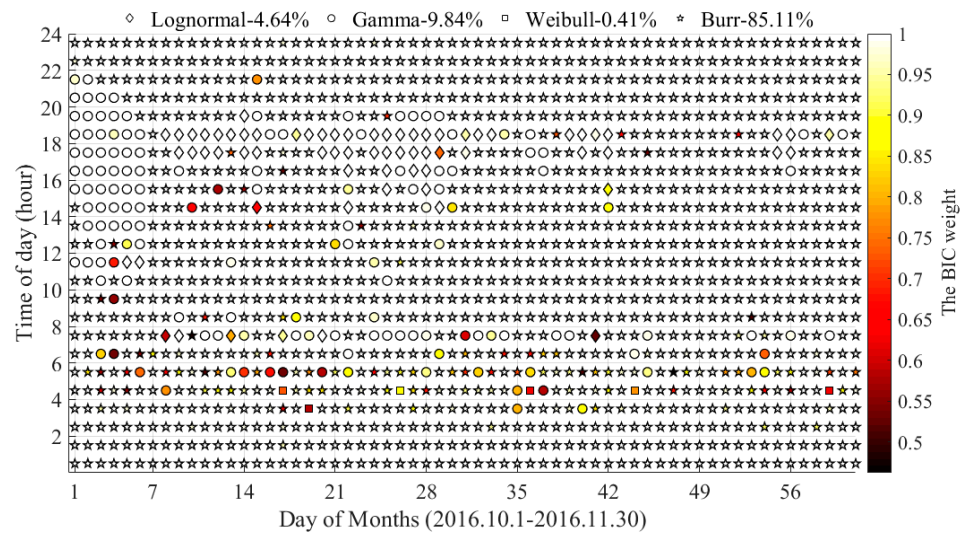


Figure 8. The fitting distributions of daily travel speed.

#### 4.5. General Distribution Selection

Based on the above analysis, the Gamma distribution is dominated in the fitting of travel distance and travel time, while the Burr distribution is more suitable for travel speed. Trips following other distributions only makes up a very small portion of the total trips. Now, it may be questioned whether it is possible to fit all the corresponding data with the dominant distribution separately? If so, how much worse will the fit be? To answer this question, the K–S test, the BIC difference, mean absolute error (MAE), and mean absolute percentage error (MAPE) are analyzed separately.

As shown in Table 5, among the trip metrics, 101, 348, and 218 out of 1,464 1 h intervals are, respectively, replaced by Gamma or Burr distribution. For travel distance, the K–S test considers both the alternative distribution (Gamma) and the best-fit distribution as acceptable, respectively, for about 90% (93 or 90 out of 101) of the data. Meanwhile, the alternative distribution performs better for travel time because more data (76%) pass the K–S test. However, for the travel speed metric, the opposite is true, which needs to be further explained by other indicators. Moreover, selection of the more complex model indicates that the benefit of improved model fit outweighs the cost of added model complexity.

Table 5. Fitting indicators of alternative distributions for three trip metrics.

Indicators		Travel Distance	Travel Time	Travel Speed
Alternative distribution	Type	Gamma	Gamma	Burr
	Number	101	348	218
K–S test	Best-fit	90 (89%)	46 (13%)	137 (63%)
	Alternative	93 (92%)	264 (76%)	0
The BIC difference	$BIC_{alternative} - BIC_{min}$	4.92	9.88	47.13
	$\Delta/BIC_{min}$	0.20%	0.10%	0.15%
MAPE (10th, 50th, 90th)		2.35%	2.51%	0.96%
MAE (10th, 50th, 90th)		0.10	0.24	0.17

For the first two trip metrics, the BIC difference between the alternative distribution and the best-fit distribution is relatively small, both being less than 10. However, the BIC difference for travel speed is slightly larger, possibly due to the different magnitude of the BIC values between the metrics. The ratio ( $\Delta/BIC_{min}$ ) of BIC difference to the BIC of the best-fit distribution is less than 0.5%, which also indicates that there is no significant difference between the two distributions from model selection based on the BIC. In addition, the MAPE and MAE of the fitted distribution and sample distribution at the 10th, 50th, and 90th percentiles are calculated by comprehensively considering the mean and variance.



A MAPE of less than 4% further demonstrates the feasibility of fitting all data with a dominant distribution.

## 5. Discussion

Based on the analysis in the previous section, the direct statistics (i.e., travel distance, and travel time) of hourly trips all follow Gamma distribution, while the indirect statistic (i.e., travel speed) obeys Burr distribution. Table 6 lists the average indicators of the trip metrics, which reflects the performance of the Gamma and Burr distributions. The bigger the  $R^2$  is, the better the goodness-of-fit is. Alternatively, smaller MAE, MAPE, and PPC also indicate a better goodness-of-fit. Three key percentiles, 10th, 50th, and 90th, were adopted to calculate the MAE and MAPE. The confidence level is equal to 80% (i.e.,  $\alpha = 0.2$ ,  $90\% - 10\% = 80\%$ ) was adopted to construct the confidence interval.

**Table 6.** Evaluation of general distribution for three trip metrics.

Indicators	Travel Distance	Travel Time	Travel Speed
General distribution	Gamma	Gamma	Burr
$R^2$	0.9859	0.9915	0.9864
MAPE (10th, 50th, 90th)	1.66%	2.15%	0.89%
MAE (10th, 50th, 90th)	0.0537	0.1887	0.2097
POPI	20.85%	20.00%	20.12%
POOI	19.21%	19.95%	19.86%
PPC	4.31%	2.54%	3.11%

According to Table 6, the mean  $R^2$  values of three trip metrics exceed 0.98, and even reach 0.9915 for travel time. This shows that the fitting distribution has a strong ability to interpret data, and this model is also good at fitting data. In general, a higher  $R^2$  indicates a stronger interpretation ability of the fitting model to the data; that is, a better fitting effect. Meanwhile, the MAPE and MAE indicators further prove this. The MAE of travel distance is less than 0.1km, lower than 0.2 min for travel time, and about 0.21 km/h for travel speed. Less than 3% of MAPEs further illustrate that the fitting distribution is quite consistent with the observed data.

The mean POPIs are slightly worse than the target (20%), which indicates that less than 80% of observation data are covered by the fitted confidence interval. Higher POPI means lower POOI. As mentioned earlier, the fitting distributions work best when both POPI and POOI are close to the target value (20%). The PPC values, 4.31%, 2.54%, and 3.11%, respectively, show a low deviation degree of POPI and POOI, which also declares that the fitting distributions of three trip metrics have good accuracy. In summary, these results indicate that the Gamma distribution fits direct trip metrics, such as travel distance and travel time, well, while the Burr distribution fits travel speed better.

## 6. Conclusions

This study models the distributions of human mobility metrics based on actual trajectory datasets, including about 18,000 online car-hailing rides, collected in Xi'an, China. Three trip metrics—travel distance, travel time, and travel speed—are highlighted in order to establish a base for human mobility research. Results of this study provided several new insights on relationships within human mobility.

First, the mobility metrics tend to right-skewed distribution rather than normal distribution based on online car-hailing trajectory data. By analyzing the daily and hourly trip data, five of the most widespread right-skewed distributions (i.e., Lognormal, Gamma, Weibull, Burr, and Rayleigh) in the scientific literature were selected as candidate distributions. By leveraging the Bayesian information criterion (BIC), we comprehensively analyzed the goodness of fit and complexity of the candidate distributions for each metric, thus acquiring the best fitting distribution and suitable parameters. The empirical



results based on online car-hailing trajectory dataset in Xi'an, China, have provided strong evidence that the mobility metrics obey the right-skewed distribution.

Second, the distribution types of mobility metrics vary, along with day of week and time of day, which means that a single distribution cannot fit all the daily and hourly data well. Initially, the Gamma distribution performs best among all alternative distributions for travel distance and can uniformly fit all daily data. Then, the Gamma or Burr distribution can only achieve a good fit in part of the daily travel time or speed data. For the hourly data, the best-fit distributions vary among alternative distributions, especially at night. The Gamma distribution most often performs better than the other four distributions for both travel distance and travel time, while the Burr distribution performs best for travel speed.

Third, although uncertain distribution types exist in the daily and hourly data, a dominant distribution exists in each mobility metric. For example, the Gamma distribution can fit more than 90% of hourly travel distance data, and the Burr distribution can achieve a fit for 85% of hourly travel speed data. Further analysis shows that it is feasible to fit all hourly data with a dominant distribution, respectively.

It is expected that the findings from this study can promote understanding about intra-urban human mobility and lay a solid foundation for human mobility research. However, we do note several limitations of this research. Firstly, the candidate distributions are limited to five commonly used skewed distributions, and more may need to be considered. Secondly, only the distribution of daily and hourly trip data is fitted and analyzed; the more fine-grained distribution (e.g., 30 min, 15 min) is also of interest, which still needs further study. Last but not the least, distribution may vary with the different datasets; multi-source data need to be taken into account to confirm the conclusions.

**Author Contributions:** Conceptualization, Chaoyang Shi and Qingquan Li; methodology, Chaoyang Shi and Shiwei Lu; software, Chaoyang Shi; formal analysis, Shiwei Lu and Xiping Yang; writing—original draft preparation, Chaoyang Shi; writing—review and editing, Qingquan Li and Xiping Yang. All authors have read and agreed to the published version of the manuscript.

**Funding:** The work was funded in part by the National Natural Science Foundation of China under Grant 41901390 and 41901392, in part by the Fundamental Research Funds for the Central Universities under Grant 2019kfyXJJS142 (HUST), in part by the Natural Science Foundation of Hubei Province under Grant 2019CFB098, in part by Open Research Fund of State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University (No. 19S03), and in part by Open Fund of Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources KF-2020-05-005.

**Data Availability Statement:** Publicly available datasets were analyzed in this study. These data can be found here: <https://gaia.didichuxing.com>.

**Acknowledgments:** The authors are grateful to Didi Chuxing for providing the data used in this paper. The authors would like to thank the anonymous reviewers and editors for providing valuable comments and suggestions which helped improve the manuscripts greatly.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Pan, G.; Qi, G.D.; Wu, Z.; Zhang, D.Q.; Li, S.J. Land-use classification using taxi GPS traces. *IEEE Trans. Intell. Transp. Syst.* **2013**, *14*, 113–123. [\[CrossRef\]](#)
2. Soliman, A.; Soltani, K.; Yin, J.J.; Padmanabhan, A.; Wang, S.W. Social sensing of urban land use based on analysis of twitter users' mobility patterns. *PLoS ONE* **2017**, *12*, e0181657. [\[CrossRef\]](#)
3. Jiang, B.; Yin, J.J.; Zhao, S.J. Characterizing the human mobility pattern in a large street network. *Phys. Rev. E* **2009**, *80*, 021136. [\[CrossRef\]](#)
4. Paolo, B.; Chiara, P.; Ramasco, J.J.; Michele, T.; Vittoria, C.; Alessandro, V. Human mobility networks, travel restrictions, and the global spread of 2009 H1N1 pandemic. *PLoS ONE* **2011**, *6*, e16591.
5. Wesolowski, A.; Eagle, N.; Tatem, A.J.; Smith, D.L.; Noor, A.M.; Snow, R.W.; Buckee, C.O. Quantifying the impact of human mobility on malaria. *Science* **2012**, *338*, 267–304. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Song, H.Y.; Lee, J.S. Finding a simple probability distribution for human mobile speed. *Pervasive Mob. Comput.* **2016**, *25*, 26–47. [\[CrossRef\]](#)

7. Brockmann, D.; Hufnagel, L.; Geisel, T. The Scaling Laws of Human Travel. *Nature* **2006**, *439*, 462–465. [[CrossRef](#)]
8. Rhee, I.; Shin, M.; Hong, S.; Lee, K.; Kim, S.J.; Chong, S. On the Levy-Walk Nature of Human Mobility. *IEEE ACM Trans. Netw.* **2011**, *19*, 630–643. [[CrossRef](#)]
9. González, M.C.; Hidalgo, C.A.; Barabási, A.L. Understanding individual human mobility patterns. *Nature* **2008**, *453*, 779–782. [[CrossRef](#)]
10. Zhao, K.; Musolesi, M.; Hui, P.; Rao, W.; Tarkoma, S. Explaining the power-law distribution of human mobility through transportation modality decomposition. *Sci. Rep.* **2014**, *5*, 9136. [[CrossRef](#)] [[PubMed](#)]
11. Calabrese, F.; Diao, M.; Lorenzo, G.D.; Ferreira, J.; Ratti, C. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transp. Res. Part C Emerg. Technol.* **2013**, *26*, 301–313. [[CrossRef](#)]
12. Jia, T.; Jiang, B.; Carling, K.; Bolin, M.; Ban, Y. An empirical study on human mobility and its agent-based modeling. *J. Stat. Mech. Theory Exp.* **2012**, *11*, P11024. [[CrossRef](#)]
13. Cho, E.; Myers, S.A.; Leskovec, J. Friendship and mobility: User movement in location-based social networks. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, 21–24 August 2011; p. 1082.
14. Kang, C.G.; Ma, X.J.; Tong, D.Q.; Liu, Y. Intra-urban human mobility patterns: An urban morphology perspective. *Phys. A Stat. Mech. Appl.* **2012**, *391*, 1702–1717. [[CrossRef](#)]
15. Jiang, S.X.; Guan, W.; Zhang, W.Y.; Chen, X.; Yang, L. Human mobility in space from three modes of public transportation. *Phys. A Stat. Mech. Appl.* **2017**, *483*, 227–238. [[CrossRef](#)]
16. Liang, X.; Zheng, X.D.; Lv, W.F.; Zhu, T.Y.; Xu, K. The scaling of human mobility by taxis is exponential. *Phys. A Stat. Mech. Appl.* **2012**, *391*, 2135–2144. [[CrossRef](#)]
17. Liang, X.; Zhao, J.; Dong, L.; Xu, K. Unraveling the origin of exponential law in intra-urban human mobility. *Sci. Rep.* **2013**, *3*, 2983. [[CrossRef](#)]
18. Bazzani, A.; Giorgini, B.; Rambaldi, S.; Gallotti, R.; Giovannini, L. Statistical laws in urban mobility from microscopic GPS data in the area of Florence. *J. Stat. Mech. Theory Exp.* **2010**, *2010*, P05001. [[CrossRef](#)]
19. Wang, W.J.; Pan, L.; Yuan, N.; Zhang, S.; Liu, D. A comparative analysis of intra-city human mobility by taxi. *Phys. A Stat. Mech. Appl.* **2015**, *420*, 134–147. [[CrossRef](#)]
20. Csáji, B.C.; Browet, A.; Traag, V.; Delvenne, J.C.; Huens, E.; Van Dooren, P.M.; Smoreda, Z.; Blondel, V.D. Exploring the mobility of mobile phone users. *Phys. A Stat. Mech. Appl.* **2013**, *392*, 1459–1473. [[CrossRef](#)]
21. Zhang, S.; Tang, J.J.; Wang, H.X.; Wang, Y.H.; An, S. Revealing intra-urban travel patterns and service ranges from taxi trajectories. *J. Transp. Geogr.* **2017**, *61*, 72–86. [[CrossRef](#)]
22. Kou, Z.Y.; Cai, H. Understanding bike sharing travel patterns: An analysis of trip data from eight cities. *Phys. A Stat. Mech. Appl.* **2019**, *515*, 785–797. [[CrossRef](#)]
23. Plötz, P.; Jakobsson, N.; Sprei, F. On the distribution of individual daily driving distances. *Transp. Res. Part B Methodol.* **2017**, *101*, 213–227. [[CrossRef](#)]
24. Zheng, Z.; Rasouli, S.; Timmermans, H. Two-regime pattern in human mobility: Evidence from GPS taxi trajectory data. *Geogr. Anal.* **2016**, *48*, 157–175. [[CrossRef](#)]
25. Chen, B.Y.; Yuan, H.; Li, Q.Q.; Lam, W.H.K.; Shaw, S.L. Map-matching algorithm for large-scale low-frequency floating car data. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 22–38. [[CrossRef](#)]
26. Krause, C.M.; Zhang, L. Short-term travel behavior prediction with GPS, land use, and point of interest data. *Transp. Res. Part B Methodol.* **2019**, *123*, 349–361. [[CrossRef](#)]
27. Xia, F.; Wang, J.Z.; Kong, X.J.; Wang, Z.B.; Li, J.X.; Liu, C.F. Exploring human mobility patterns in urban scenarios: A trajectory data perspective. *IEEE Commun. Mag.* **2018**, *56*, 142–149. [[CrossRef](#)]
28. Zhang, B.; Chen, S.Y.; Ma, Y.F.; Li, T.Z.; Tang, K. Analysis on spatiotemporal urban mobility based on online car-hailing data. *J. Transp. Geogr.* **2020**, *82*, 102568. [[CrossRef](#)]
29. Lin, M.; Hsu, W.J. Mining GPS data for mobility patterns: A survey. *Pervasive Mob. Comput.* **2014**, *12*, 1–16. [[CrossRef](#)]
30. Liu, Y.; Kang, C.G.; Gao, S.; Xiao, Y.; Tian, Y. Understanding intra-urban trip patterns from taxi trajectory data. *J. Geogr. Syst.* **2012**, *14*, 463–483. [[CrossRef](#)]
31. Tang, J.J.; Liu, F.; Wang, Y.H.; Wang, H. Uncovering urban human mobility from large scale taxi GPS data. *Phys. A Stat. Mech. Appl.* **2015**, *438*, 140–153. [[CrossRef](#)]
32. Taylor, M.A.P. Fosgerau’s travel time reliability ratio and the burr distribution. *Transp. Res. Part B Methodol.* **2017**, *97*, 50–63. [[CrossRef](#)]
33. Clauset, A.; Shalizi, C.R.; Newman, M.E.J. Power-law distributions in empirical data. *SIAM Rev.* **2007**, *51*, 661–703. [[CrossRef](#)]
34. Johnson, J.; Omland, K. Model selection in ecology and evolution. *Trends Ecol. Evol.* **2004**, *19*, 101–108. [[CrossRef](#)]
35. Zhang, K.P.; Ning, J.; Zheng, L.; Liu, Z.J. A novel generative adversarial network for estimation of trip travel time distribution with trajectory data. *Transp. Res. Part C Emerg. Technol.* **2019**, *108*, 223–244. [[CrossRef](#)]
36. Posada, D.; Buckley, T.R. Model selection and model averaging in phylogenetics: Advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Syst. Biol.* **2004**, *53*, 793–808. [[CrossRef](#)] [[PubMed](#)]
37. Akaike, H. Information measures and model selection. *Int. Stat. Inst.* **1983**, *50*, 277–291.

- 
38. Shi, C.Y.; Chen, B.Y.; Lam, W.H.K.; Li, Q.Q. Heterogeneous data fusion method to estimate travel time distributions in congested road networks. *Sensors* **2017**, *17*, 2822. [[CrossRef](#)]
  39. Shi, C.Y.; Chen, B.Y.; Li, Q.Q. Estimation of travel time distributions in urban road networks using low-frequency floating car data. *ISPRS Int. Geo Inf.* **2017**, *6*, 253. [[CrossRef](#)]