Journal of Urban Development and Management
https://www.acadlore.com/journals/JUDM

Identification Method for Unlicensed Taxis in Urban Areas: A Case Study

Rongqiao Li[1], Yun Xiao[2*]

[1] School of Traffic and Transportation, Northeast Forestry University, 150040 Harbin, China
[2] School of Urban Construction and Transportation, Hefei University, 230606 Hefei, China

[*] Correspondence: Yun Xiao (xiaoyun@hfuu.edu.cn)

Received: 02-01-2023 **Revised:** 03-04-2023 **Accepted:** 03-10-2023

Citation: R. Q. Li and Y. Xiao, "Identification method for unlicensed taxis in urban areas: A case study," *J. Urban Dev. Manag.*, vol. 2, no. 1, pp. 22-33, 2023. https://doi.org/10.56578/judm020103.

Abstract: To accurately identify unlicensed taxis, this study measured their mileage using a traffic surveillance bayonet and obtained a threshold value by fitting a function to the mileage of previously identified unlicensed taxis. Abnormal driving vehicles were identified as those with a mileage exceeding the threshold value. Through a "white list" screening process, information on suspected unlicensed taxis was obtained. An empirical analysis of City A in Anhui Province showed an identification threshold of 85.8 km for unlicensed taxis. The study identified 68 highly suspected unlicensed taxis, 513 moderately suspected unlicensed taxis, and 1595 generally suspected unlicensed taxis. Suspected unlicensed taxis had a strong correlation with taxi mileage (r=0.895, sig(2-tailed)=0), with a mean mileage of 128.5 km and standard deviation of 50.8. This mileage was less than the average taxi mileage but significantly higher than the mileage traveled by private cars (mean=25.1 km, SD=16.4). The study's contribution lies in its development of a method for accurately identifying unlicensed taxis, which has significant implications for improving transportation safety.

Keywords: Traffic engineering; Unlicensed taxis; Traffic surveillance bayonet; Vehicle mileage; Data mining

1. Introduction

Unlicensed taxis are private vehicles that engage in transportation operations without a license, posing a significant safety risk to the field of transportation. In China, the number of drivers engaged in illegal transportation operations remains high due to a lack of operational ability training and qualification audits. Unlicensed taxi activities harm passengers in various ways, such as overcharging customers by taking unnecessary detours, insufficient liability insurance coverage in case of accidents, and some drivers having records of drug driving or other crimes, posing a significant threat to passenger safety. Furthermore, most unlicensed taxis do not undergo regular safety and technical inspections, increasing traffic safety hazards.

Currently, the identification of unlicensed taxis relies mainly on manual methods such as fixed-point detection, regional searches, and passenger reporting, which consume significant human and material resources and disrupt traffic. Improving this identification method has become a matter of widespread concern in the community. Some scholars have studied the formation mechanism of illegal operations, with Ma [1] proposing a governance path for illegal operations based on a system analysis method. Lin and Wang [2] proposed a reporting system for unlicensed taxis using near-field communication technology and Android to encourage passengers to report unlicensed taxis and create an effective governance environment. However, most of the research has focused on policy, law enforcement, and regulatory levels, with limited attention given to the technical path of identifying unlicensed taxis.

Researchers have conducted simulation experiments to study the operating characteristics of unlicensed taxis. Wang et al. [3] used a convolutional neural network to obtain trajectory data on unlicensed taxis and normal vehicles, performing feature learning and recognition to improve the recognition rate of unlicensed taxis. Shuai et al. [4] proposed an unlicensed taxi identification algorithm based on k-medoids and utilizing radio frequency identification (RFID) data, which was authenticated through experiments. Ma et al. [5] collected vehicle operation data using RFID technology and established a mathematical model for identifying unlicensed taxis using a Self-

Organizing Maps (SOM) neural network clustering algorithm, which was found to be effective. While these studies provide valuable references, their theoretical validation relies heavily on data simulation, and their practical applicability requires further testing.

Scholars have employed data analysis methods to identify unlicensed taxis, leveraging the development of big data technology. Zhao et al. [6] analyzed vehicle refueling data, extracting the temporal and spatial characteristics of refueling and identifying abnormal refueled vehicles as suspected unlicensed taxis. Li et al. [7] utilized motor vehicle electronic registration identifier data to construct a detection model for unlicensed taxis through an integrated learning method. Yuan et al. [8] established a detection model to identify coarse-grained unlicensed taxis by extracting vehicle travel data from traffic surveillance bayonets, and further used the feature training support vector machine classification model to identify fine-grained unlicensed taxis. Wang et al. [9] estimated and predicted passenger preferences for public transit vehicles and unlicensed taxis using a multi-logistic model, based on passenger transportation travel demand.

While the above-mentioned studies provide valuable data analysis methods, their data sources are relatively simplistic since most unlicensed taxis do not refuel at gas stations. Therefore, analyzing illegal operating behavior through refueling data is not comprehensive enough. This paper aims to improve the extensiveness of the data by analyzing the driving characteristics of unlicensed taxis using driving data obtained from traffic surveillance bayonets. By proposing the use of driving mileage as a critical index for detecting and identifying unlicensed taxis and determining a driving mileage threshold value, the paper aims to enhance the scientific and practical applicability of the identification method.

Researchers both domestically and internationally have utilized big data analysis of traffic surveillance bayonets and vehicle satellite positioning information to extract abnormal vehicle operation trajectories for further study. Huang et al. [10] utilized a self-coding network and long-term and short-term memory neural network to extract and reconstruct vehicle tracks, upon which they built an abnormal track recognition model. Their model demonstrated a 9.8% higher F1 value than support vector machine, random forest, and long-term and short-term memory neural network models on average, effectively identifying abnormal vehicle tracks. Xu et al. [11] proposed a dual-mode detection model to analyze abnormal vehicle motion behavior. They obtained vehicle information from the background model and fused the results of the dual model analysis to ultimately determine abnormal vehicle motion behavior. The results showed a relatively low detection error for their model. Athanesious et al. [12] presented a vehicle trajectory detection model based on trajectory clustering algorithms, using spectral clustering theory and Bayesian clustering theory to define events and determine whether the vehicle trajectory was abnormal. Their model demonstrated a 12% higher accuracy than existing technology. Chawla et al. [13] analyzed potential abnormal routes of vehicle operation by constructing an OD matrix, and identified accurate abnormal road sections and nodes based on historical information of vehicle operation. Lei [14] proposed a maritime framework anomaly trajectory detection model (MT-MAD), which segmented trajectories, obtained trajectory feature scores within the region, and sorted them before determining abnormal trajectories by comparing them with the threshold of trajectory feature scores. The experimental results showed that this model can effectively detect maritime anomaly trajectories. Zhang et al. [15] utilized the spline function model to calculate the ship's trajectory, and their results showed that the prediction accuracy of this model was better compared to linear regression model, polynomial regression model, and weighted regression model. Biskas and Babu [16] used Hidden Markov Model (HMM) to determine the probability of abnormal vehicle operation based on the local trajectory of vehicle motion and compared it with the threshold to determine the abnormal vehicle operation trajectory. Overall, these studies utilized function fitting, constructing travel matrices, and machine learning algorithms to construct vehicle operation models that can effectively identify abnormal vehicle operation trajectories.

Traffic surveillance bayonets are widely distributed and time-sensitive devices that provide high-quality data on vehicle travel characteristics. As a result, many scholars have conducted related research based on data obtained from these devices [17-20]. This paper aims to analyze the driving characteristics of unlicensed taxis using driving data obtained from traffic surveillance bayonets. By using driving mileage as a critical index for detecting and identifying unlicensed taxis, and determining a driving mileage threshold value to flag vehicles that exceed this value as suspected unlicensed taxis, this paper aims to improve the accuracy and effectiveness of identifying unlicensed taxis. By utilizing the data from traffic surveillance bayonets in this way, the paper aims to contribute to the development of a more comprehensive and practical method for identifying unlicensed taxis.

2. Research Method

2.1 Step Identification

This paper collected all vehicle driving data through traffic surveillance bayonets, calculated the vehicle driving mileage (S), and compared it with the driving mileage threshold (S')to identify unlicensed taxis. Therefore, we built 6 datasets, as shown in Figure 1 below.

First, the traffic surveillance adjacent bayonet distance matrix is established through traffic surveillance information (TSI) datasets, and is matched with passing vehicle information (PVI) datasets to calculate the average daily mileage (S_q) of each vehicle q.

Second, the mileage data of the identified unlicensed taxis in the unlicensed taxi information (UTI) datasets are calculated, and S' is determined as the threshold value for identifying unlicensed taxis.

Third, a vehicle q is added to the abnormal driving vehicle information (ADVI) datasets when its driving mileage is greater than or equal to S'.

Finally, the abnormal driving vehicle information (ADVI) datasets are compared with the compliance vehicle information (CVI) datasets, and after eliminating the compliant vehicle information, the suspected unlicensed taxi information is obtained and categorized in the suspected unlicensed taxi information (SUTI) datasets.

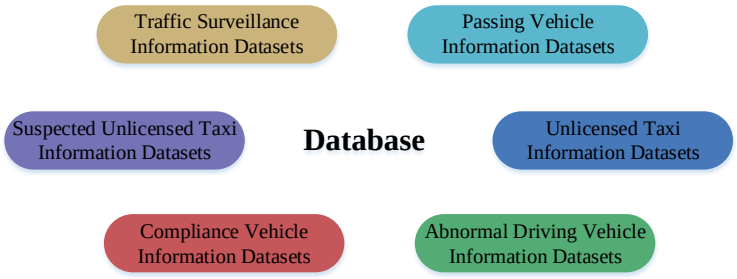

Figure 1. Detection and identification database of suspected unlicensed taxis

2.2 Calculation of Vehicle Mileage Data

2.2.1 Establishing the bayonet distance matrix

A traffic surveillance bayonet is a road traffic site detection system at a road intersection that captures images, identifies, and records the vehicles that pass through it. Its basic attributes include the bayonet number (tscID), name (tscName), longitude (longitude), and latitude (latitude). When storing the traffic surveillance bayonet attribute information to the TSI dataset, the expression is as follows:

$$TSI = (tscID_i, tscName_i, longitude_i, latitude_i) \tag{1}$$

Related studies have shown that obtaining vehicle trajectories through traffic surveillance bayonet location information and road nodes can yield better results [21, 22]. In this paper, the distance between adjacent bayonets tsci and tscj is obtained based on traffic surveillance bayonet information in the TSI dataset, and the vehicle miles traveled are calculated. The road section between two adjacent bayonets tsci and tscj is divided into several key nodes N_1, N_2, ..., N_i, which correspond to the latitude and longitude coordinates of (x_1,y_1), (x_2,y_2), ..., (x_i,y_i), respectively. The distance between two adjacent bayonets can be expressed by the following equation:

$$D_{i,j} = \sum_{1}^{i} R * arcos[cos(y_{m-1}) * cos(y_m) * cos(x_{m-1} - x_m) + sin(y_{m-1}) * sin(y_m)] \tag{2}$$

where, R is the radius of the earth, 6371.0 km, and m is the m^{th} critical node on the road section between two adjacent bayonets, $2 \le m \le i$.

A traffic surveillance adjacent bayonet distance matrix (n) is established as shown in Table 1 below, where "l_{ij}" refers to a distance in meters between adjacent traffic surveillance bayonets i and j along the road section:

Table 1. Traffic surveillance distance matrix of adjacent bayonets

Number	TSI_1	TSI_2	…	TSI_i	TSI_j	…	TSI_n
TSI_1	0	l_{12}	…	l_{1i}	l_{1j}	…	l_{1n}
TSI_2	l_{21}	0	…	l_{2i}	l_{2j}	…	l_{2n}
…	…	…	…	…	…	…	…
TSI_i	l_{i1}	l_{i2}	…	0	l_{ij}	…	l_{in}
TSI_j	l_{j1}	l_{j2}	…	l_{ji}	0	…	l_{jn}
…	…	…	…	…	…	…	…
TSI_n	l_{n1}	l_{n2}	…	l_{n3}	l_{n3}	…	0

2.2.2 Data collection of passing vehicles

When a vehicle passes a road intersection, the traffic surveillance bayonet collects and records relevant data on the vehicle, including the passing number (*pID*), vehicle number plate (*vnp*), *tscID*, passing time (*pTime*), among others. The information on the vehicle passing the traffic surveillance bayonet can be expressed by the following equation:

$$r = (pID, vnp, tscID, pTime) \tag{3}$$

The passing vehicle data collected by the traffic surveillance bayonet during the statistical period T is stored in the PVI dataset, as shown in Eq. (4) below, and the results are shown in Table 2.

$$R = \{ r_1, r_2, \cdots, r_n \} \tag{4}$$

where, r_i=(pID_i, vnp_i, $tscID_i$, $pTime_i$).

Table 2. Information on vehicles passing through the bayonet

Passing number	Vehicle number plate	Bayonet number	Passing time
1832927287	Wan A****6	34060300001190150005	2021/12/1 6:59
1832927701	Wan F****5	34060300001190150005	2021/12/1 6:59
1832927436	Wan F****9	34060300001190150005	2021/12/1 6:59
…	…	…	…
1832928383	Wan F****2	34060300001190110049	2021/12/1 6:59

2.2.3 Vehicle track data cleaning

The crossing data r_q of a vehicle q in the PVI dataset is matched with the traffic surveillance bayonet attribute information in the TSI dataset according to the sequence of crossing time, and the calculation of the single driving mileage of vehicle q is completed as shown in the following equation:

$$s_q = \sum_{1}^{i} D_{i-1,i} \tag{5}$$

where, $D_{i,j}$ is the distance between bayonet i and j, km.

For all trips traveled by vehicle q during the statistical period T, the cleaning process is completed, and its average daily mileage S_q is calculated as follows:

$$S_q = \frac{\sum_{1}^{i} s_m}{T} \tag{6}$$

where, S_q is the average daily mileage, km, S_m is the vehicle m^{th} trip mileage, $1 \le m \le i$, and T is the unlicensed taxi identification statistics period, days.

2.3 Indicator Identification Threshold

In the UTI dataset, the mileage data of identified unlicensed taxis are extracted, and the recognition index threshold S' is calculated. Firstly, basic information on unlicensed taxis is filtered out from the UTI dataset and compared with the PVI dataset to calculate the average daily mileage S'_j of identified unlicensed taxis during the statistical period T:

$$S'_j = \sum_{1}^{i} \frac{l_j}{iT} \tag{7}$$

where, S'_j is the average daily mileage of the *jth* vehicle in the statistical period T, l_j is the mileage of the j^{th} unlicensed taxis within the statistical period T, km, $1 \le j \le i$, and i is the total number of identified unlicensed taxis in the city.

Function fitting has been used by previous scholars to determine study thresholds [23-25]. In this paper, this approach is referenced to establish the relationship between unlicensed taxi mileage and the proportion of unlicensed taxis. The best-fit function *f(x)* is obtained through function fitting, and the mileage threshold is selected

as the average daily mileage when the percentage of unlicensed taxis is N%, based on the function characteristics. The travel mileage threshold can be expressed by the following equation:

$$S' = f^{-1}(N) \tag{8}$$

2.4 Determining the Suspected Unlicensed Taxis

The average daily vehicle mileage *Sq* during the statistical period T is compared with the identification index threshold *S′*. If the following Eq. (9) is satisfied, the vehicle q is considered an abnormal driving vehicle and will be added to the ADVI dataset:

$$S_q \geq S' \tag{9}$$

Taxis, compliant online vehicles, and police cars in cities are known to accumulate large mileages during their work, and are therefore defined as compliant vehicles. These vehicles are deposited into the CVI dataset, matched with the ADVI dataset, and eliminated from the ADVI dataset. Information on suspected unlicensed taxis is obtained and deposited into the SUTI dataset as shown in the following equation:

$$SUTI = ADVI - CVI \tag{10}$$

3. Empirical Analysis

3.1 Research Subjects

City A is a Type II large city located in the northeast region of Anhui Province, covering a total area of 2,741 square kilometers and having a resident population of 1,970,300. The city is situated on the border of Henan and Anhui provinces and has a significant passenger flow, providing a market space for unlicensed taxis. For this study, a dataset consisting of 1,627,000 vehicle travel records collected by traffic surveillance bayonets in City A from October 1 to October 5 was used, as shown in Table 2.

3.2 Calculation of Vehicle Miles Traveled

3.2.1 Bayonet distance matrix

The four fields of the 1,095 traffic surveillance bayonets in city A (tscID, name tscName, longitude, and latitude) are stored in the TSI dataset, as shown in Table 3 below.

Table 3. Basic attribute information from traffic surveillance bayonets in City A

Bayonet number	Bayonet name	Longitude	Latitude
34062100001190410137	XX County Hospital East XC-516	116.761	33.915
34062100001190410170	LQ-517, Lulou Village, a town in XX County	116.687	33.913
34060400001190450018	XX County Wutong Avenue and Riverside Road South	116.843	33.840
…	…	…	…
34060300001190150006	XX County Victory Avenue and Magnolia Avenue	116.765	33.903

Using the TSI dataset, the distance of the road sections between adjacent traffic surveillance bayonets is calculated according to Eq. (2) above. Based on this calculation, a distance matrix of the traffic surveillance bayonets in City A is established, as shown in Table 4 below.

Table 4. City A traffic surveillance bayonet distance matrix (unit: meters)

Number	TSI_1	TSI_2	TSI_3	…	TSI_{1095}
TSI_1	0	122.7	0	…	0
TSI_2	122.7	0	120.6	…	0
TSI_3	0	120.6	0	…	130.7
…	…	…	…	…	…
TSI_{1095}	0	0	130.7	…	0

3.2.2 Specific measurement of mileage

The vehicle travel data in the PVI dataset is matched with the bayonet attribute information in the TSI dataset to obtain the vehicle travel track information. The single vehicle mileage Sq and the average daily vehicle mileage Sq are calculated using Eq. (5) and (6) above. As a result, the final mileage data of 162,500 vehicles in 5 days were obtained (Mean=28.01 km, SD=30.53), and are shown in Table 5 below.

Table 5. Information on vehicle driving distance in City A in 5 days

Vehicle number plate	Total mileage(km)	Travel days (days)	Average daily mileage (km)
Wan F****0	2753.2	5	550.6
Wan F****6	2452.5	5	490.5
Wan F****3	2402.5	5	480.5
…	…	…	…
Wan A****6	20.0	4	5.00

3.3 Measurement of Identification Index Thresholds

The identification threshold was determined using a sample of 228 identified unlicensed taxis in City A. Following the method described above, the average daily mileage data (Mean=163.32 km, SD=77.21) of these 228 identified unlicensed taxis were extracted, and are shown in Table 6 below.

Table 6. Unlicensed taxi mileage

Vehicle number plate	Total mileage(km)	Travel days (days)	Average daily mileage (km)
Wan F****9	1151.4	5	322.4
Wan F****6	1074.4	5	300.8
Wan F****9	1060.7	5	297.0
…	…	…	…
Wan F****5	94.7	5	26.5

The Extreme function, Gauss function, Logistic function, and Lorentz function were used to fit the data on the average daily mileage of the identified unlicensed taxis. The relationship between the mileage of the identified unlicensed taxis and the proportion of identified unlicensed taxis was obtained and is shown in Figure 2. The results of the data fitting comparison are presented in Table 7. Based on the table, the average daily mileage data of the identified unlicensed taxis best fit the Logistic function (AIC=49.12, BIC=54.91).

Figure 2 shows that the relationship between the proportion of unlicensed taxis and their average daily mileage is not a simple linear decrease. Different average daily mileages have a greater impact on the proportion of unlicensed taxis. Using the Logistic function, this study found that the proportion of unlicensed taxis decreases more smoothly when the driving distance is below 100 km, and more rapidly when the driving distance is between 100 km and 200 km. To avoid missing more unlicensed taxis and misclassifying private cars as suspected unlicensed taxis, this study selects the average daily driving mileage at which the percentage of unlicensed taxis is 85% as the driving mileage recognition threshold, which is 85.8 km.

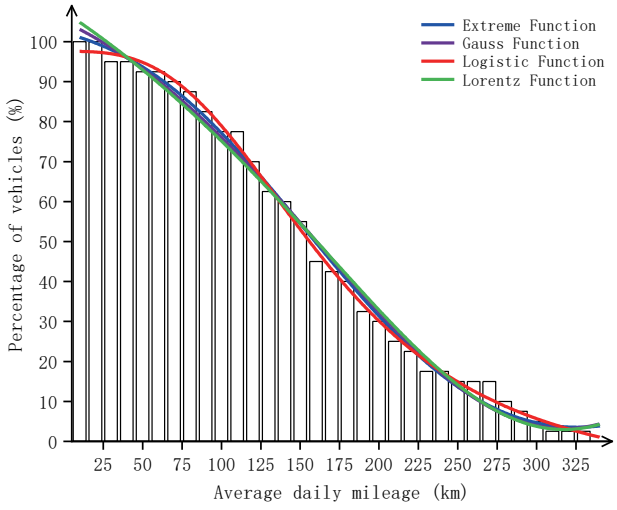

Figure 2. Function fitting results

Table 7. Data fitting comparison results

Function types	Function	AIC	BIC
Extreme	$double \quad z = \frac{x - x_c}{w}$ $y = y_0 + A * \exp(-\exp(-z)) - z + 1$	70.79	76.28
Gauss	$y = y_0 + \frac{A}{w * sqrt(pi/2)} * \exp(-2 * (\frac{x - x_c}{w})\hat{}2)$	83.99	89.48
Logistic	$y = A_2 + \frac{A_1 - A_2}{1 + (x/x_0)\hat{}2}$	49.12	54.91
Lorentz	$y = y_0 + 2 * \frac{A}{pi} * (\frac{w}{4 * (x - x_c)\hat{}2} + w\hat{}2)$	91.81	97.30

To verify the reasonableness of this threshold, the average rate of change of the fitted Logistic function was calculated, which reflects the sharpness of the vehicle occupancy with changes in driving mileage. The average rate of change of the Logistic function is 0.17 when the average daily mileage of unlicensed taxis is between 0 and 85.8 km, and 0.50 when the average daily mileage of unlicensed taxis is between 85.8 km and 171.6 km. These results indicate that the average rate of change is small when the mileage is between 0 and 85.8 km, and the proportion of unlicensed taxis decreases slowly with an increase in the average daily mileage. However, when the average daily mileage is between 85.8 km and 171.6 km, the average rate of change becomes significantly larger, and the proportion of unlicensed taxis decreases more quickly with an increase in the average daily mileage. Therefore, it is more reasonable to choose 85.8 km as the threshold for mileage identification, as it is at the point where the function's average rate of change varies significantly and the proportion of unlicensed taxis decreases with an increase in the average daily mileage.

3.4 Identification of Unlicensed Taxis

3.4.1 Identification of vehicles exceeding the driving threshold

If the average daily mileage S_q of a vehicle q within 5 days is greater than or equal to the recognition threshold S', the vehicle q is considered an abnormal driving vehicle and is added to the ADVI dataset. The final information on the average daily mileage of 4,007 abnormal driving vehicles (Mean=174.53 km, SD=83.73) was obtained and is shown in Table 8 below.

Table 8. Abnormal driving vehicle information

Vehicle number plate	Average daily mileage (km)
Wan F****0	550.6
Wan F****6	490.5
Wan F****3	480.5
…	…
Wan F****6	85.8

3.4.2 Identification of suspected unlicensed taxis

When identifying suspected unlicensed taxis based on driving mileage, it is necessary to screen out driving data from compliant vehicles, taxis, police cars, and other official vehicles that were collected at the same time. The compliant vehicle information in the CVI dataset is matched with the vehicle information in the ADVI dataset to eliminate compliant vehicle information. As a result, 1,794 pieces of compliant online car and taxi driving data, 15 pieces of police car driving data, and 22 pieces of driving test vehicle driving data were screened out. This left us with a total of 2,176 suspected pieces of unlicensed taxi driving mileage data (Mean=128.5 km, SD=50.8), which were added to the SUTI dataset, as shown in Table 9 below.

Table 9. Suspected unlicensed taxi information

Vehicle number plate	Average daily mileage (km)
Wan F****0	550.6
Wan F****6	470.7
Wan F****8	422.5
…	…
Wan F****6	85.8

3.5 Analysis of the Characteristics of Suspected Unlicensed Taxis

3.5.1 Analysis of distribution

(1) Mileage characteristics

To explore the characteristics of suspected unlicensed taxis in terms of mileage, the mileage data of the 2,176 suspected unlicensed taxis were divided into 5 intervals, as shown in Figure 3. During the five-day period from October 1 to October 5, the proportion of suspected unlicensed taxis with an average daily mileage below 150 km was larger, with 41.77% having an average daily mileage between 100 km and 150 km, and 35.29% having an average daily mileage below 100 km. When the average daily mileage exceeded 200 km, the proportion of suspected unlicensed taxis was relatively small, accounting for only 8.37%.

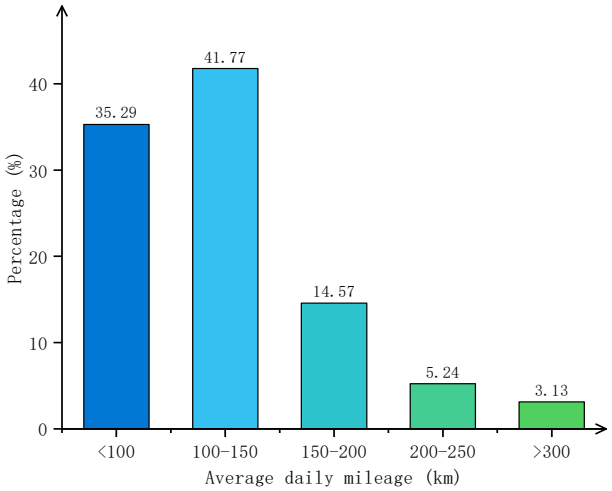

Figure 3. The average daily mileage of suspected unlicensed taxis

(2) Vehicle distribution characteristics

The majority of the identified suspected unlicensed taxis are located in City A, but there are also some vehicles in neighboring cities. To explore the distribution characteristics of the suspected unlicensed taxis, a classification was made according to the attribution of vehicle number plates, and the results are shown in Figure 4 below.

Out of all the suspected unlicensed taxis, 1,618 vehicles were located in City A of Anhui Province, accounting for 74.36% of the total. 272 vehicles were located in City B, Anhui Province, accounting for 12.50%. 58 vehicles were located in City C, Anhui Province, accounting for 2.67%. 25 vehicles were located in City D, Henan Province, accounting for 1.15%. 49 vehicles were located in City E, Jiangsu Province, accounting for 2.25%. There were also 154 vehicles located in other areas, accounting for 7.08% of the total. The presence of suspected unlicensed taxis from other cities accounted for 25.64%, which can be attributed to the fact that City A is located at the intersection of two multi-city provinces and has a high density of passenger traffic.

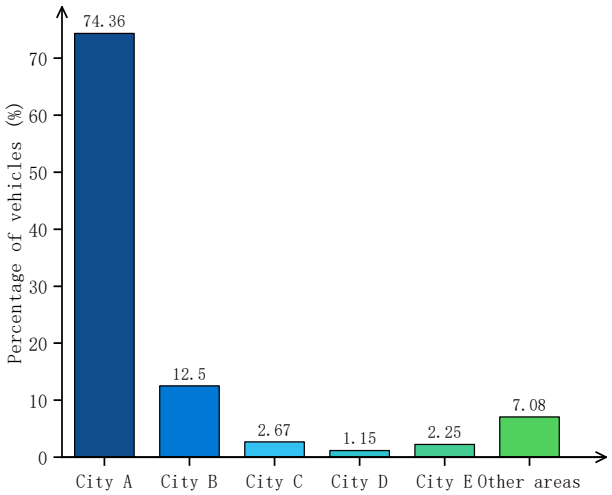

Figure 4. Regional distribution of suspected unlicensed taxis

(3) Classification of suspected unlicensed taxis

To explore the possibility that the 2,176 vehicles in the suspected SUTI dataset are engaged in illegal operations, a K-mean cluster analysis was conducted using IBM SPSS Statistics 26 with the average daily mileage parameter, and the initial number of clusters was set to 3. The final cluster centers were obtained after 18 calculation iterations, as shown in Table 10 below.

According to the clustering results, among the 2,176 suspected unlicensed taxis, 68 vehicles were highly suspected to be unlicensed taxis, accounting for 3.13%. There were also 513 vehicles that were moderately suspected to be unlicensed taxis, accounting for 25.58%, and 1,595 were generally suspected to be unlicensed taxis, accounting for 73.3%, as shown in Figure 5 below. These results further improve the accuracy of detecting and identifying unlicensed taxis, which can facilitate a more reasonable allocation of attention when regulators identify unlicensed taxis.

Table 10. Final clustering center situation

Form	General	Moderately	Highly
Average daily mileage clustering centroids	104.4	178.3	317.1

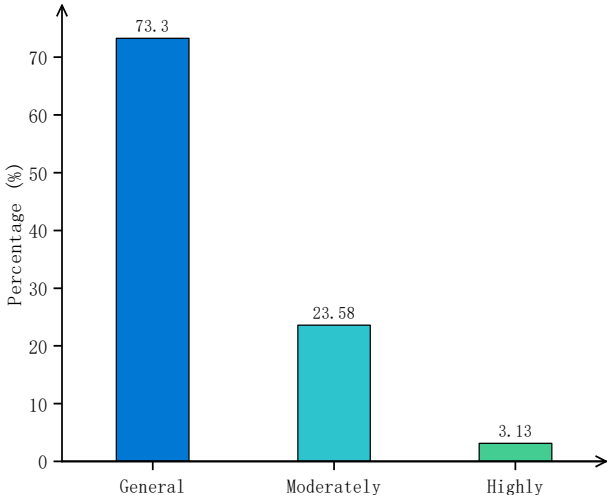

Figure 5. Suspected unlicensed taxi classification

3.5.2 Comparative analysis with taxis

To explore the similarities and differences between suspected unlicensed taxis and licensed taxis in City A in terms of mileage, driving data on 1,563 taxis (Mean=239.0 km, SD=84.1) were extracted for the purpose of conducting a correlation test. The results are shown in Table 11 below.

According to the results, it can be seen that there is a significant and strong correlation (r=0.895) between the average daily mileage of suspected unlicensed taxis and licensed taxis. This suggests that there may be similarities in the driving patterns of suspected unlicensed taxis and licensed taxis in City A.

Table 11. Correlation test between suspected unlicensed taxis and licensed taxis

		Suspected unlicensed taxis	Taxis
suspected unlicensed taxis	Pearson Correlation	1	.895
	Sig.(2-tailed)		0
taxis	Pearson Correlation	.895	1
	Sig.(2-tailed)	0	

The mileages of suspected unlicensed taxis and licensed taxis were divided into 8 intervals, as shown in Figure 6 below. It can be seen that the average daily mileage of suspected unlicensed taxis is generally distributed within the same range as the average daily mileage of licensed taxis, with peak ranges of 100 km to 150 km and 200 km to 250 km, respectively. After exceeding the peak range, the average daily mileage of suspected unlicensed taxis and licensed taxis decreases in their respective mileage ranges. However, when the average daily mileage falls below the 100 km interval, the average daily mileage of suspected unlicensed taxis is significantly greater than the average daily mileage of licensed taxis, accounting for 35.29% and 4.54%, respectively.

The operating efficiency of licensed taxis is generally higher than that of suspected unlicensed taxis. This may

be due to several reasons. Firstly, licensed taxis have distinctive signs, drivers have undergone rigorous training, and vehicles have passed regular inspections, which increases the level of trust that the general public has in them. Secondly, taxis operate more efficiently as they can travel on roads to find passengers on the one hand, and accept taxi calls from passengers through online taxi platforms on the other. In contrast, unlicensed taxis need to be aware of traffic enforcement officers while looking for passengers, and are thus less likely to travel to areas where traffic police are more vigilant [26].

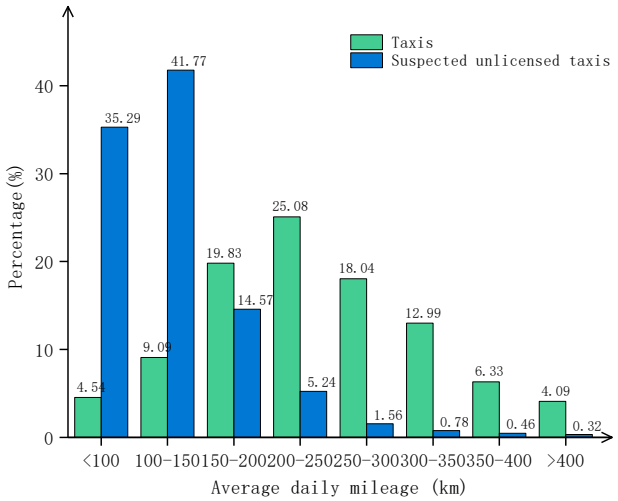

Figure 6. Mileage comparison of suspected unlicensed taxis and licensed taxis

3.5.3 Comparative analysis with private cars

A correlation test was conducted between the mileage of suspected unlicensed taxis and the mileage of private cars (Mean=25.1 km, SD=16.4), and the results are shown in Table 12.

It can be seen that there is only a weak correlation (r=0.111) between the average daily mileage of suspected unlicensed taxis and the average daily mileage of private cars, indicating that there is a significant difference between the two groups in terms of mileage.

Table 12. Correlation test between suspected unlicensed taxis and private cars

		Suspected unlicensed taxis	**Private cars**
suspected unlicensed taxis	Pearson Correlation	1	0.111
	Sig.(2-tailed)		0
private cars	Pearson Correlation	0.111	1
	Sig.(2-tailed)	0	

4. Conclusion

This study investigated the detection and identification of unlicensed taxis using big data, combining traffic surveillance bayonet data and vehicle driving data to obtain vehicle driving mileage information. By applying a function fitting method to determine the driving mileage recognition threshold, this study effectively identified unlicensed taxis and achieved strong applicability. Through the example of City A in Anhui Province, 2,176 suspected unlicensed taxis (Mean=128.5 km, SD=50.8) were ultimately identified using driving mileage as the recognition index.

The results of this study showed that there is a significant and strong correlation between the mileage of suspected unlicensed taxis and licensed taxis in City A (Mean=239.0 km, r=0.895, sig(2-tailed)=0). However, the mileage of unlicensed taxis is generally smaller than that of licensed taxis. In addition, there is a significant difference between the mileage of suspected unlicensed taxis and private cars in City A (Mean=25.1 km, r=0.111, sig(2-tailed)=0), with the mileage of suspected unlicensed taxis being approximately 9.52 times that of private cars.

Further research on unlicensed taxis can be conducted in several aspects. Firstly, this study used mileage as the main indicator to identify unlicensed taxis, which may omit occasional illegal operational behaviors. Therefore, future research could explore the establishment of an illegally operating passenger vehicle identification system based on multiple indicators, such as mileage and the number of passing cars. Secondly, this study did not propose control measures for unlicensed taxis, and a next step might be to explore and study methods for the accurate

seizure and punishment of unlicensed taxis based on spatio-temporal travel data, as well as the preventive supervision of illegal operating driver groups.

In conclusion, this study provides valuable insights into the detection and identification of unlicensed taxis using big data, and provides a foundation for future research on the control and management of unlicensed taxis.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] H. S. Ma, "Research on the formation mechanism and improvement path of illegal carriages in the transportation of Shenzhen," *J. Transp. Eng. Inf.*, no. 2, pp. 64-69, 2012.

[2] J. Y. Lin and L. X. Wang, "Design and implementation of illegal operating vehicles' reporting system based on NFC and android," *Tech. Autom. Appl.*, vol. 37, no. 8, pp. 119-122, 2018.

[3] N. Wang, P. F. Duan, and L. L. Wu, "Research of illegal operation vehicle identification based on convolutional neural network," *J. Comput. Appl.*, vol. 36, no. S2, pp. 193-196, 2016.

[4] D. Shuai, Z. L. Lan, and Y. C. Li, "Recognising illegal operation vehicles based on k-mediods and its improved algorithm," *Comp. Appl. Softw.*, vol. 33, no. 5, pp. 154-157, 2016. http://dx.chinadoi.cn/10.3969/j.issn.1000-386x.2016.05.038.

[5] S. Y. Ma, C. D. Sun, and J. L. Zhu, "Application of SOM neural network in the identification of illegal operating vehicles," *Chin. Comput. Commun.*, no. 8, pp. 83-85, 2017.

[6] Q. H. Zhao, T. H. Jiang, F. Zhao, and B. Ma, "Abnormal vehicle detection based on spatio-temporal big data," *Transd. Microsyst. Technol.*, vol. 38, no. 4, pp. 139-142, 2019.

[7] L. Chen, L. J. Zheng, L. Xia, W. N. Liu, and D. H. Sun, "Detecting and analyzing unlicensed taxis: A case study of Chongqing City," *Phys. A: Statist Mech. Appl.*, vol. 584, Article ID: 126324, 2021. https://doi.org/10.1016/j.physa.2021.126324.

[8] W. Yuan, P. Deng, T. Taleb, J. F. Wan, and C. F. Bi, "An unlicensed taxi identification model based on big data analysis. *IEEE Trans. Intell Transp. Syst.*, vol. 17, no. 6, pp. 1703-1713, 2016. https://doi.org/10.1109/TITS.2015.2498180.

[9] B. Y. Wang, L. Y. Zhao, Y. Q. Pang, D. Zhang, and X. G. Yang, "Analysis of passenger's choice between shuttle bus and illegal taxi," *Pro. Soc Behav Sci.*, vol. 96, pp. 1948-1960, 2013. https://doi.org/10.1016/j.sbspro.2013.08.220.

[10] S. C. Huang, C. F. Shao, J. Li, X. Y. Zhang, and J. P. Qian, "Vehicle trajectory reconstruction and anomaly detection using deep learning," *J. Transp. Syst. Eng. Inf. Technol.*, vol. 21, no. 3, pp. 47-54, 2021. https://doi.org/10.16097/j.cnki.1009-6744.2021.03.006.

[11] Y. Xu, X. Ouyang, Y. Cheng, S. N. Yu, L. Xiong, C. C. Ng, S. Pranata, S. M. Shen, and J. L. Xing, "Dual-mode vehicle motion pattern learning for high performance road traffic anomaly detection," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, December 16, 2018, IEEE, pp. 145-152. https://doi.org/10.1109/CVPRW.2018.00027.

[12] J. J. Athanesious, S. S. Chakkaravarthy, S. Vasuhi, and V. Vaidehi, "Trajectory based abnormal event detection in video traffic surveillance using general potential data field with spectral clustering," *Multimed Tools Appl.*, vol. 78, pp. 19877-19903, 2019. https://doi.org/10.1007/s11042-019-7332-y.

[13] S. Chawla, Y. Zeng, and J. Hu, "Inferring the root cause in road traffic anomalies," In 2012 IEEE 12th International Conference on Data Mining, Brussels, Belgium, December 10-13, 2012, IEEE, pp. 141-150. https://doi.org/10.1109/ICDM.2012.104.

[14] P. Lei, "A framework for anomaly detection in maritime trajectory behavior," *Knowl. Inf. Syst.*, vol. 47, pp. 189-214, 2016. https://doi.org/10.1007/s10115-015-0845-4.

[15] L. Y. Zhang, Q. Meng, Z. Xiao, and X. J. Fu, "A novel ship trajectory reconstruction approach using AIS data. *Ocean Eng.*, vol. 159, pp. 165-174, 2018. https://doi.org/10.1016/j.oceaneng.2018.03.085.

[16] S. Biswas and R. V. Babu, "Anomaly detection via short local trajectories," *Neurocomputing*, vol. 242, pp. 63-72, 2017. https://doi.org/10.1016/j.neucom.2017.02.058.

[17] A. Crouzil, L. Khoudour, P. Valiere, and D.N.T. Cong, "Automatic vehicle counting system for traffic monitoring," *J. Electron. Imaging*, vol. 25, no. 5, Article ID: 051207, 2016. https://doi.org/10.1117/1.JEI.25.5.051207.

[18] C. Y. Fang, K. B. Jia, and P. Y. Liu, "Identification of taxi violation behavior based on surveillance video,"

Comput. Simul., vol. 37, no. 5, pp. 326-331, 2020.
[19] M. Fernández-Sanjurjo, B. Bosquet, M. Mucientes, and V. M. Brea, "Real-time visual detection and tracking system for traffic monitoring," *Eng. Appl. Artif. Intel.*, vol. 85, pp. 410-420, 2019. https://doi.org/10.1016/j.engappai.2019.07.005.
[20] Y. Malinovskiy, Y. J. Wu, and Y. H. Wang, "Video-based vehicle detection and tracking using spatiotemporal maps," *Transport. Res. Rec.*, vol. 2121, no. 1, pp. 81-89, 2009. https://doi.org/10.3141/2121-09.
[21] Y. T. Luo, W. Tao, M. N. Yang, and Y. K. Zhang, "Historical driving track set based visual vehicle behavior analytic method," *Comput. Sci.*, vol. 48, no. 9, pp. 86-94, 2021. https://doi.org/10.11896/jsjkx.200900040.
[22] T. Li, Y. K. Zhu, X. H. Wu, Y. P. Xiao, and H. F. Wu, "Vehicle trajectory prediction method based on intersection context and deep belief network," *J. Electron. Inf. Technol.*, vol. 43, no. 5, pp. 1323-1330, 2021.
[23] Y. L. Ma, S. M. Qi, H. T. Wu, and L. Y. Fan, "Traffic conflict identification model based on post encroachment time algorithm in ramp merging area," *J. Transp. Syst. Eng. Inf. Technol.*, vol. 18, no. 2, pp. 142-148, 2018. http://dx.doi.org/10.16097/j.cnki.1009-6744.2018.02.022.
[24] P. P. Wu, W. Y. Cai, G. D. Tang, Z. Q. Wang, and Z. F. Zhu, "Laser range measuring system based on dynamic multi-threshold error correction method," *J. Electron. Meas. Instrum.*, vol. 35, no. 7, pp. 170-177, 2021.
[25] R. S. Ba, J. Li, X. D. Zhou, G. B. Zhen, H. L. Xu, L. Ding, Y. J. Li, and J. Na, "Uncertainty analysis of 1 on 1 laser induced damage threshold measurement," *Acta Metrol. Sin.*, vol. 43, no. 1, pp. 26-34, 2022.
[26] T. L. M., "Can the legalization of online car-hailing resolve the "black car"," *People's Trib.*, no. S1, pp. 60-62, 2016.