Article

Predicting the Place Visited of Floating Car: A Three-Layer Framework Using Spatiotemporal Probability

Wenwen He ¹ and Fu Ren ^{1,2,*}

¹ School of Resource and Environmental Sciences, Wuhan University, Wuhan 430079, China; hewenwen@whu.edu.cn

² Key Laboratory of Geographic Information Systems, Ministry of Education, Wuhan University, Wuhan 430074, China

* Correspondence: renfu@whu.edu.cn

Abstract: Human-flow pattern can reflect the urban population mobility and the urban operating state. Understanding the trajectory of urban-population moving patterns can improve the effectiveness of urban-management measures. While most of the existing studies on human moving have placed a huge emphasis on location forecasting through the types of activities humans take part in and urban land-use types, this type of forecasting research is limited to relying on specific activity types and land-use types. The urban-population moving pattern has spatial and temporal characteristics, and this feature greatly affects the prediction of where humans will visit. This study aimed to predict the possible places to visit by using the spatiotemporal model. We analyzed the itinerary characteristics of urban taxis and proposed a model based on the taxi itinerary characteristics to predict the drop-off locations. This model can be used to predict the possible arrival locations of urban taxis. We selected three grids of travel data from each period in another day to test the prediction accuracy of the proposed model. The results show that the model can predict the destination of urban taxis to a certain degree.

Keywords: human flows; taxi trips; spatiotemporal moving patterns; location visited prediction

Citation: He, W.; Ren, F. Predicting the Place Visited of Floating Car: A Three-Layer Framework Using Spatiotemporal Probability. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 663. <https://doi.org/10.3390/ijgi10100663>

Academic Editor: Wolfgang Kainz

Received: 16 August 2021

Accepted: 28 September 2021

Published: 1 October 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 widespread deployment of sensors provides a way for the collection of big data, and massive amounts of big data provide basic data resources for the mining of deep information. In the urban environment, useful information mined from human-flow datasets covering all aspects of the city provides researchers with key information and decision support for scientifically planning urban functional areas, rationally dispatching urban resources, and effectively responding to emergencies. The acquisition of massive amounts of human movement data enables city managers to analyze historical activity information and use historical information to predict the direction of future urban activities, so as to rationally allocate urban resources and promote efficient urban operations.

Applying different research models to a variety of activity trajectory datasets can dig out a variety of potential information about urban operations and provide scientific support from multiple perspectives for promoting urban development. For example, in recent years, many studies have used social-media data, mobile-phone data, sign-in data, subway-card data, etc., to understand the population distribution and population flow of the city, thereby discovering hot spots in the city [1–4]. Combining land-use data with GPS data for taxi-demand analysis and hot-spot detection provides a reference for taxi resource allocation [5–7]. We use public-transportation-trajectory data and smart-card data to identify major public-transportation corridors in order to increase the utilization rate on limited road resources [8,9].

Research based on massive taxi-trajectory data provides the possibility to manage urban traffic and monitor human activities [10]. The research based on taxi GPS trajectory

data can be summarized as mining the driving pattern and trajectory data to analyze the traffic pattern in the city [11,12], estimating the travel demand and travel mode according to the traffic state, and then evaluating the urban road traffic conditions [10,13]. Assisting the operation and management of taxis by studying the behaviors of taxi drivers, such as seeking passengers and driving patterns, helps improve the efficiency of taxi operations [14–16].

Although relevant research has been able to mine useful information from the massive amount of historical taxi-trajectory data, the research on the prediction of future-activity information is relatively scarce. Most of the existing research focuses on the exploration and analysis of the temporal and spatial characteristics of urban population activities and is dedicated to discovering the temporal and spatial hot spots of these activities [17–19]. Tang et al. proposed a probabilistic model based on the Hidden Markov Model (HMM) to predict the travel path of taxi drivers [20]. Zheng and Zhou applied the scaling-law method to study the dynamic spatial access frequency of taxi trajectory data and proposed a model to predict urban time and space arrivals from points of interest (POIs) [21].

The only research related to the prediction of passenger's visit location by taxi mainly focuses on inferring the passenger's possible visit function area from the passenger's boarding location and boarding time [22], as well as the activities that may be engaged in after arriving at the destination. There is a lack of the utilization of history trajectory data to predict the passenger's visit to the destination unit. Gong et al. considered space and time constraints, constructed a Bayesian-rule-based access probability model for points of interest, and combined it with Monte Carlo simulation to study Shanghai taxi trajectory data [23].

On the one hand, studying the areas that urban residents may visit can assist in the rational allocation of public transportation resources, make full use of the limited urban transportation resources, and optimize people's travel patterns. On the other hand, it helps city managers to grasp the overall situation of the city's activity space and manage the city's daily operation more effectively. The purpose of this article is to analyze the historical operating-trajectory data of urban floating vehicles, construct a spatiotemporal probability model based on historical spatiotemporal data, and predict that urban residents may visit the destination unit by taxi. In this study, we propose a three-layer framework, using the spatiotemporal probability (TI-STPM) model to predict the user's purpose of travel. We used the time of pickup and drop-off, the location of pickup and drop-off, the distance of travel as historical data to build our model; what is more, we took road-network and bus-line data as auxiliary factors to participate in the model calculation. Finally, the likely drop-off location was predicted according to the time and location the passenger was picked up. The research results may help fill in the application of temporal and spatial probability models in urban public transportation and provide a reference for the study of prediction of place visit probability.

2. Methodology

The flowchart of the proposed TI-STPM is illustrated in Figure 1. The procedure includes three parts: (1) The abnormal data of the original taxi trajectory data are filtered and analyzed of the temporal characteristics of taxi data. (2) K-means clustering and kernel density are applied to analyze the temporal and spatial distribution characteristics of taxi data, and the study area is divided into hexagonal grid cells of suitable size with hexagonal grid according to the analysis results. (3) Based on the number of boarding and alighting, the location of the boarding and alighting, the time of getting on and off, the travel distance, the road networks, the bus lines, and the divided hexagonal grid unit, a multilayer spatiotemporal probability model is established.

Figure 1. Workflow of the location visited prediction, using the spatiotemporal probability model.

2.1. Data Preprocessing and Spatiotemporal Analysis

2.1.1. Dataset and Data Filtering

In this study, we selected a dataset that contains about 1,145,562 taxi travel records from five workdays in the week of 3 June to 7 June (Monday to Friday) 2019 as the research dataset. In the week which we selected, there was only light rain on Monday and cloudy weather from Tuesday to Friday, with temperatures ranging from 25 to 31 degrees Celsius on all five weekdays, this kind of weather is a little hot, so it will increase the demand and possibility of people taking taxis. The data were collected within Xiamen, China (all administrative districts except Gulangyu Island). As shown in Figure 2, located on the southeast coast of China, Xiamen is one of the special economic zones specially approved by the State Council of China. The city has 6 administrative regions, with a total area of 1700.61 km². By the end of 2019, the number of taxis in Xiamen was about 6572; the main areas of taxi operation were concentrated the in districts of Siming, Huli, Haicang, and Jimei; and 80% of the city's taxis operated on Xiamen Island.

To filter the datasets, we deleted the abnormal records caused by positioning errors, transfer errors, or operation errors, such as pickup and drop-off locations out of our study area or the coordinates value is zero, the travel distance less than 1500 m or more than 30 km, etc. After filtering out these records, there were 1,089,957 records left, and the properties are shown in Table 1; each record contains information regarding the car number, pickup date, pickup location (longitude and latitude), drop-off date, drop-off location (longitude and latitude), and pass mile.

Figure 2. Location of the study area.

Table 1. Sample records of travel data.

Car Num	Pickup Date	Pickup_Lon	Pickup_Lat	Drop-Off_Date	Drop-Off_Lon	Drop-Off_Lat	Pass Mile
8027f4gh	6/05 8:34	118.178853	24.521353	6/05 8:58	118.149598	24.533922	8.6
a71c64ac	6/05 9:48	118.101252	24.469193	6/05 10:06	118.101252	24.469193	12.7

2.1.2. Temporal Characteristic Analysis

We found from the statistical results of the travel numbers that, in five days (Figure 3), the peak hours of boarding occurred at 9:00 a.m., 3:00 p.m., and 11:00 p.m., while at 5:00 a.m., 1:00 p.m., and 5:00 p.m., the boarding numbers are at a relatively low value. The peak hours of alighting occurred at 10:00 a.m., 3:00 p.m., and 11:00 p.m., while the relatively low values also appeared at 1:00 p.m., 5:00 a.m., and 6:00 p.m.

2.2. Clustering Using *k*-Means and Kernel Density with Hexagon

2.2.1. K-Means Cluster Analysis

After analyzing the data for temporal characteristics, we used the *k*-means method and kernel density method for clustering analysis.

Clustering is a classification technique that aims at partitioning a dataset into clusters such that the objects within a cluster are similar and the objects in different clusters are dissimilar according to certain predefined criteria [24,25]. The clustering algorithms can be summarized as partitioning methods, hierarchical methods, density-based methods, grid-based methods, model-based methods, etc. The *k*-means clustering algorithm is a widely used partitioning method in many study areas [26]. The *k*-means clustering algorithm clusters the data according to the characteristics of the data themselves, without artificial labeling; therefore, in this study, we applied the *k*-means method to cluster the two variables of boarding location and travel distance. In addition, because the *k* value is a hyperparameter and it generally needs to be selected by experience, it was important to choose the *k* value.

Figure 3. Hourly variation of drop-off (a) and pickup (b) number (24-h clock).

In the research of applying k-means cluster analysis, there are many methods for determining the value of k , and the elbow method is a heuristic method for determining the number of clusters in a dataset. The method plots the explained change as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use [27]. Thus, in our study, we applied the elbow method to help determine a reasonable k value, and Figure 4 illustrates the variation of the cluster deviation of boarding location and travel distance at various periods of the day with the value of k . It can be seen that, when the k value is 8, we could get the best classification effect. The results of the k-means clustering analysis of the travel distance and boarding location are shown in Figure 5.

Figure 4. Best k-value chosen by elbow method in different periods.

Figure 5. Cont.

Figure 5. Cluster analysis of the travel distance and boarding location for each period. (a–l) rerepresent 12 periods spectively in a day.

Figure 5 shows the clustering results of boarding location and travel distance for each period in a day. From Figure 5, it can be seen that the minimum distance clusters for residents to travel by taxis is no less than 2.4 km and no more than 25 km, and the cluster of minimum travel distance is between 10:00 p.m. and 12:00 p.m., while the maximum travel distance is between 6:00 p.m. and 10:00 p.m. Among them, the cluster of minimum travel distance (2.4~7 km) is mainly concentrated in Xiamen Island, and the distribution of travel distance in the range 7.5~13 km is mainly concentrated in the northwest area outside Xiamen Island. As shown in Figure 5f–k, besides concentrated distribution in Xiamen Island, the travel distance of 8~10 km is mainly distributed in the northern area outside Xiamen Island (Class 4), and the time is concentrated between 10:00 a.m. and 10:00 p.m. The reason may be that this area is the location of Xiamen North High-Speed Railway Station, and the minimum distance from this place to Xiamen Island is about 9 km, indicating that passengers departing from this area during this period prefer to take a taxi to the Xiamen Island area. The clusters with a travel distance of about 13 km are mainly concentrated in the central and southwestern areas of Xiamen Island, from 12:00 a.m.–8:00 a.m. (Figure 5a,b,d). Starting from 8:00 a.m., the area with a distance of about 17 km is mainly concentrated in the northeast area of Xiamen Island, and it lasts until 10:00 p.m. (Figure 5e–k). It can be seen from Figure 5c,d that, when the travel distance is about 13 km, there are obvious clustering characteristics in the northwest area outside Xiamen Island from 4:00 a.m. to 10:00 a.m. From Figure 5h–k, it can be seen that, when the travel distance is about 16~18 km, the trips are mainly concentrated in the northeast area of Xiamen Island and at the period of 4:00 a.m.–10:00 p.m.

2.2.2. Kernel Density Analysis

After using k-means cluster analysis on travel number, we also performed kernel density analysis on travel number to obtain hot spots for boarding and alighting. The results of the kernel density analysis on the pickup and drop-off points are shown in Figure 6 below. It can be seen from the result that the boarding and alighting hot spots are mainly clustering in the central business district on Xiamen Island. In addition, in the north area of Xiamen Island, near the Xiamen North High-speed Railway Station, is also a high-density area of boarding and alighting area.

Figure 6. Kernel density analysis of the pickup number (a) and drop-off number (b).

2.2.3. Divide the Study Area with Hexagons

According to the results of k-means clustering analysis and kernel density analysis, we used a hexagonal grid to divide the study area. There are many ways to divide the geographical space, while the space division requires graphics to cover the space completely, neither leaving the area nor overlapping coverage. There are regular triangles, regular squares, and regular hexagons that can divide a spaceplane without intersecting each other.

The reason why this paper chose regular hexagons (hexagonal grid) for spatial division includes two reasons. First, when the side lengths are equal, the regular hexagon has the largest area. The advantage is that, in the division process, the same number of graphics has the largest division area, and the coverage rate is the highest. Second, the distance between one centroid and any of the six neighboring centroids is the same, thus reducing the sampling bias compared to the square grid [28]. When using regular hexagons to divide the study area, we tested regular hexagons with side lengths of 100, 300, 500, and 700 m. We assumed that any two points between two adjacent polygons can be reached directly, and we found that, when a regular hexagon has a side length of 300 m, the maximum distance between any two points inside two adjacent polygons is 1960 m, which is basically in line with the minimum travel distance we chose to take a taxi (>1500 m). Therefore, we used a regular hexagonal grid with a side length of 300 m to divide the study area. The result of the division is shown in Figure 7 below.

Figure 7. Using hexagons to divide the research area: (a) whole study area and (b) example area of the study area.

2.3. Three-Layer Framework, Using Spatiotemporal Probability

The structure of the three-layer framework is shown in Figure 8b. A probability factor is calculated for each layer. after calculating the P_{i_num} , P_{i_vol} , and P_{i_time} , the visit probability of the grid i $P_{i_drop-off}$ is as follows:

$$P_{i_drop-off} = \prod_{i=1}^N P_{i_num} P_{i_vol} P_{i_time} R_{i_density} B'_{i_density} \quad (1)$$

Figure 8. (a) Schematic diagram of the grid where the pickup point location. (b) Schematic diagram of three-layer framework. L1: the net inflow ratio. L2: the maximum possible visit distance of the grid i . L3: the probability of visit in a specific period of the grid.

In the first layer (Figure 8b(L1)), we used the ratio of the number of cars dropped off to the number of cars picked up in a grid as the net inflow ratio (P_{i_num}). The output of this layer was used to divide the grid into two types, with values of '0' and '1'. As Figure 9 shows below, we counted the total inflow (blue line) travel numbers as drop-off volume and the total outflow (red line) travel numbers as pickup volume in each hexagonal grid and divided the total drop-off volume by the total pickup volume. This can be expressed as Equations (2)–(5). If the result value is less than 1, it means that the grid is a net outflow unit; otherwise, the grid is regarded as a net inflow unit. If the drop-off number is 0, then the value of P_{i_num} is 0; if the pickup number is 0, then the value of P_{i_num} is equal to the drop-off numbers. The calculation results are shown in Table 2. We normalized the comparison value result and took P_{i_num} as [0, 1]. When the $\frac{N_{i_drop}}{N_{i_pick}}$ value is less than 1, the P_{i_num} value is 0; otherwise, is 1. Then we assigned the P_{i_num} value to the hexagonal grid.

$$P_{i_num} = \frac{N_{i_drop}}{N_{i_pick}} \quad (2)$$

$$N_{i_drop} = \sum_{i=1}^n n_{i_drop} \quad (3)$$

$$N_{i_pick} = \sum_{i=1}^n n_{i_pick} \quad (4)$$

$$P_{i_num} = \begin{cases} 1, & N_{i_pick} \leq N_{i_drop} \\ 0, & N_{i_pick} > N_{i_drop} \end{cases} \quad (5)$$

where P_{i_num} denotes the net inflow ratio, N_{i_drop} denotes the drop-off volume in grid i , and N_{i_pick} denotes the pickup volume in grid i .

Figure 9. Inflow and outflow of a grid.

Table 2. Part of the calculated results of the model.

FID	P_{i_num}	$P_{i_vol} \cdot 10^3$ (%)	P_{i_time}	ATD (km)	$R_{i_density} \cdot 10^3$	$B'_{i_density} \cdot 10$	$P_{i_drop-off} \cdot 10^6$
0	0.285714	0.080274	0.88888	18.80	0.000062	0.068027	0.0859849
1	0.214285	1.677728	0.36594	18.13	0.002092	0.002043	0.5622808
2	0.133333	1.079652	0.38624	15.96	0.001658	0.002578	0.2376545
3	8.812500	0.042414	0.35547	11.92	0.000334	0.012771	0.5667384
4	0.894736	0.140836	0.52173	14.13	0.000181	0.023584	0.2806405
5	0.421052	3.675701	0.33569	9.69	0.004291	0.000996	2.2204049
6	6.789473	0.035582	0.37908	25.00	0.000183	0.023255	0.3897308
7	0.421052	7.280948	0.36435	8.21	0.000865	0.004943	4.7758333
8	39.16666	4.332417	0.35599	13.43	0.001496	0.002857	258.18227
9	0.038461	75.91045	0.34937	5.91	0.007248	0.000590	4.3619225
10	0.285714	0.080274	0.88888	18.80	0.001060	0.004032	0.0871317
...	...	...	...	...	...	...	...
153	1147.4540	0.066895	0.44615	14.55	0.000490	0.008710	146.15849

In the second layer (Figure 8(L2)), the travel distance was used as the input data. We calculated the percentage of the total travel distance of all drop-off points in grid i to the total travel distance of all drop-off points in the whole study area as the maximum possible visit distance of the grid i (P_{i_vol}). In this layer, the travel distance was used as a calculating factor of the visit probability of a grid; the greater the total distance of all drop-off trips in the grid, the greater the probability that the grid will be visited. For example, in the case where the travel distance in Layer2 is d , according to the result of the probability of getting off in Layer1, in the neighborhood grid around the pickup grid A1, the grids B1 and B2 are both possible drop-off grids, but according to the result of the maximum probability of drop-off in Layer2, grid B1 has a higher visiting probability than grid B2, so grid B1 is more likely to be the visiting location. The calculated probability of drop-off in Layer2 can be expressed as Equations (6)–(8).

$$P_{i_vol} = \frac{V_{i_drop}}{V_A} \quad (6)$$

$$V_{i_drop} = \sum_{i=1}^n d_{drop} \quad (7)$$

$$V_A = \sum_{i=1}^n \sum_{j=1}^m v_{j_drop} \quad (8)$$

where P_{i_vol} denotes the ratio of grid i to the total travel distance in the study area, V_{i_drop} denotes the total travel distance of all drop-offs points in grid i , and V_A denotes the total travel distance of all drop-off points in the whole study area. The calculated results are shown in Table 2.

In the third layer, we divided the trip data by hour and counted the first three periods with the largest number of drop-offs in each grid, and we calculated the percentage of the number of drop-offs in these three time periods to the total number of drop-offs in the whole study area as the probability of visit in a specific period of the grid (P_{i_time}).

Urban residents usually visit a specific place at a specific time; for example, they go to the company's location at work time and to the restaurant's location at lunchtime. Therefore, in this layer, we considered the probability of time visits in the grid. Firstly, we calculated the number of drop-offs per hour in each grid. Secondly, we took the period of the first three hours with the largest number of drop-offs as the possible access time of the grid. Finally, we calculated the percentage of the total number of drop-offs in the first three time periods to the total number of all drop-offs in the study area and took it as the visit probability of the grid, which is P_{i_time} , as shown in Figure 10. According to the calculation results of this layer, with the travel distance as the visit radius, it is possible to infer the locations that may be visited within the travel range at a certain time. The P_{i_time} can be expressed as Equations (9)–(12).

$$P_{i_time} = \frac{N_{i_t}}{N_T} \quad (9)$$

$$N_{i_t} = N_{i_t1} + N_{i_t2} + N_{i_t3} \quad (10)$$

$$t1, t2, t3 \in (\text{Max}\{2, 4, 6, \dots, 24\}, 3) \quad (11)$$

$$N_T = \sum_{i=1}^3 \sum_{j=1}^m N_{ij} \quad (12)$$

where P_{i_time} denotes the time visit probability of the grid i ; N_{i_t} denotes the total number of the first three hours with the largest number of drop-offs of the grid i ; N_{i_t1} , N_{i_t2} , and N_{i_t3} denote the number of drop-offs in the first, the second, and the third periods, respectively; and N_T denotes the total number of all drop-offs in the study area.

Figure 10. Period with the largest number of drop-offs.

The visit probability of a place includes two aspects: temporal and spatial. After the calculation of Layer1, Layer2, and Layer3 above, each grid has the attribute of temporal and spatial visit probability. As a supplement to the spatiotemporal probability model, we calculated the average travel distance (ATD) of the grid unit, which is defined as follows: the travel distance of a passenger to a certain place by vehicle. The calculation method is the sum of the distances of all drop-off points in grid i divided by the number of drop-offs in grid i , as shown in Equations (13)–(15). The ATD is used to search for the getting off grid within the radius of d when the travel distance d is known.

$$D_{i_mean} = \frac{V_{i_drop}}{N_{i_drop}} \quad (13)$$

$$V_{i_drop} = \sum_{j=1}^n d_j \quad (14)$$

$$N_{i_drop} = \sum_{i=1}^n n_{drop} \quad (15)$$

where D_{i_mean} denotes the ATD, V_{i_drop} denotes the sum of the distances of all drop-off points in grid i , and N_{i_drop} denotes the number of drop-offs in grid i .

Road network is one of the common complex network systems. The operation of urban floating vehicles is a behavior restricted by the urban road network, and the traffic travel of the urban population is largely affected by the structure of the road network. In the study of urban population traveling, using the statistical characteristics of the urban road network as one of the influencing factors can well describe the status of the human flow under different road network structures. Therefore, in this study, we took the urban road network data as one of the calculation factors of TI-STPM and hoped to improve the accuracy of the proposed prediction model.

The density of the road network usually represents the accessibility of an area. The higher the road network density, the higher the accessibility of the area, which represents a higher visited probability in this article. Therefore, in the calculation of the road network factor, we used the road network density of a grid ($R_{i_density}$) as the calculation factor. To calculate the road network density of a grid, we first applied the regular hexagon grid constructed above to cut the road network in the study area and count the total length of all roads in each grid. Then we took the ratio of the total length of roads in each grid to the

area of the regular hexagonal grid as the density of road network in a grid. The calculation method is shown in Equation (16).

$$R_{i_density} = \frac{d_{i_road}}{A_i} \quad (16)$$

where $R_{i_density}$ denotes the density of road network in grid i , d_{i_road} denotes the sum of the distance of road network in grid i , and A_i denotes the area of grid i .

The urban bus is one of the important parts of urban public transportation. In daily urban transportation trips, on the one hand, buses are used as a taxi connection to participate in the human flows; on the other hand, buses are used as alternative transportation to taxis, and there is a competitive relationship between buses and taxis. The degree of development of bus lines affects the service number of taxis in the area. Generally, the more developed bus lines will lead to lower taxi-service numbers, and passengers are less likely to visit the area by taxi. When the bus participates in travels as a city's transport connection, it is usually the mode of transportation that passengers choose before taking a taxi; that is to say, people will choose to take the bus first, and then take the taxi (after finishing bus ride) to visit the place where the bus cannot reach. Therefore, in this study, we just considered the second situation; that is, buses affect the service of urban taxis in a competitive relationship.

In this article, we defined $B_{i_density}$ as the density of bus lines, that is, the total length of bus lines per grid area. Since buses have an adverse effect on the service number of taxis, we took the reciprocal of the bus-line density, that is, $B'_{i_density} = \frac{1}{B_{i_density}}$, as an influencing factor to calculate the probability of a place visited. Similar to calculate the density of the road network above, the calculating process of the $B_{i_density}$ in a grid was as follows: We first applied the hexagonal grid to cut the bus lines in the study area and counted the total length of all bus lines in each grid. Then we took the ratio of the total length of bus lines in each grid to the area of the regular hexagonal grid as the density of bus line in a grid. Finally, we took the reciprocal of the calculation result ($B_{i_density}$) as a calculating factor ($B'_{i_density}$) of the proposed model. The calculation method is shown in Equations (17) and (18).

$$B_{i_density} = \frac{d_{i_busline}}{A_i} \quad (17)$$

$$B'_{i_density} = \frac{A_i}{d_{i_busline}} \quad (18)$$

where $B_{i_density}$ denotes the density of bus lines in grid i , $B'_{i_density}$ denotes the reciprocal of the bus line density, $d_{i_busline}$ denotes the sum of the distance of bus lines in grid i , and A_i denotes the area of grid i .

Finally, we assigned the calculated ATD of drop-offs in each grid to the corresponding grid, used it as the basis to determine the distance of a single trip, and combined it with the three-layer spatiotemporal probability model to predict the visited location of the inputted boarding time and boarding location.

3. Results

In this part, we analyze the results of the proposed method. The results include two parts: the calculation results of the TI-STPM and the accuracy of the model testing.

3.1. The Calculation Results of TI-STPM

In the first part of the conclusion, we used 869,985 taxi travel records (80% of the dataset) to perform model calculations. Table 2 provides more details of the calculation results. Due to the large length of the calculation results, we only list some of the results for display. The first column of Table 2 represents the ID value of the hexagonal grid; the second to fourth columns are the intermediate results of the model calculation; and the last column is the final result of the model, that is, the visit probability of each grid. Figure 11a,b,d

corresponds to the second, third, and fourth columns of Table 2, respectively, and Figure 11c represents the average travel distance of all trips in each hexagon. Figure 11h corresponds to the last column of Table 2.

Figure 11. (a,b) Calculated result of the L1 and L2. (c) ATD of each hexagonal grid. (d) Calculated result of L3. (e) Distribution of the number of trips in each study area in the first three hours. (f) Density of road net. (g) Reciprocal of bus line density. (h) Final result of the model.

It can be seen from Table 2 and Figure 11a the distribution of the inflow and outflow ratio (P_{i_num}) in each hexagonal grid. Among all the hexagons, about one-third (50 of 154) of them have negative net inflows, thus showing that these areas are more likely to be used as origin areas for taxi trips, while the remaining two-thirds (104 of 154) are net inflow areas; that is, places within these areas are more destinations for passengers. Moreover, there are a considerable number of hexagonal grid areas with large net inflow ratios, indicating that these areas have an absolute advantage in being visited over the other areas which have a low net inflow ratio. It also can be seen from Figure 11a that these areas with a high probability of drop-off are mainly concentrated in the north and eastern area outside of Xiamen Island (Figure 11a red hexagons), and the other areas with high visited probability are distributed in the western area of Xiamen Island (Figure 11a orange hexagons).

It can be found from the calculation result of P_{i_vol} (Figure 11b) that the travel distance distribution of each hexagonal grid unit is relatively concentrated, and the calculation result is similar to the calculation result of P_{i_num} . However, in this layer (Layer2), the hexagons have a higher probability of being visited and are more concentrated, and there is a clear dividing line from the low-value hexagons. It can be seen that the high-value regions are more concentrated on Xiamen Island, except for a high-value region in the northern region of the study area. Figure 11c (D_{i_mean}) shows that the ATD in Xiamen Island has the shortest ATD, while the north of the study area shows a longer ATD. Based on the results of Figure 11b, it can be seen that areas with a shorter ATD have the higher visit probability, and the area with a longer ATD has a lower visit probability.

Figure 11d shows the calculation results of the Layer 3 of the model. In this layer, we analyzed the time distribution of trips in each hexagonal grid unit. The results show that, in the north of the study area, the first three hours with the largest number of trips occupy most of the trips of the entire hexagonal grid, indicating that there are obvious travel time clustering characteristics of trips in these areas. Even though there are higher values in the south and west of the study area, we speculate that the number of trips in these areas is relatively small and the travel time is relatively single, which leads to the higher P_{i_time} values in these areas, but at the same time, this result indicates that these areas do not have obvious clustering characteristics of travel time. Figure 11e as the intermediate result of P_{i_time} calculation shows the distribution of the number of trips in each study area in the first three hours.

The calculation results of the $R_{i_density}$ and $B'_{i_density}$ are shown in Figure 11f,g. From Figure 11f, we can see that the areas with the highest road network density are mainly in the northwestern area of Xiamen Island, as the main commercial area of Xiamen Island; the development of the road network in this area is higher than in other areas; and other areas with high road network density are distributed in the northeast area (Huli District) of Xiamen Island, which is the main residential area in Xiamen Island. In addition, the western area outside of Xiamen Island (Haicang District) and the northern area outside of Xiamen Island (Jimei District) also have high road network density. These two areas are developed administrative districts, except for Xiamen Island, so they also have a high road network density.

From the calculation results of $B'_{i_density}$ (Figure 11g), it can be seen that the areas with low bus-line density are mainly distributed in areas outside Xiamen Island, especially in the edge areas of Xiamen city. In these areas, due to the lack of bus lines, there is usually a higher demand for taxi services, and residents are more likely to travel by taxi. While in the Xiamen Island, the western areas outside of Xiamen Island and the northern areas outside of Xiamen Island have higher bus line density, indicating that these areas have higher bus line service levels, which could decrease the service number of taxis. Thus, after analyzing the calculation results of the $B'_{i_density}$, we took it as a negative factor to calculate the probability of taxi visits.

From the calculation result of TL-STPM (Figure 11h), it can be seen that most of the hexagons with higher visit probability are distributed in the Xiamen Island, and the probability value is highest in the northwest part of the island. That is because this area is

the core economic zone of Xiamen Island and has a relatively high scale of human flows during the working day. The northeast area of Xiamen Island, which is the main residential area on the island, also has a high probability of visiting. In addition, as the central tourist concentration area of Xiamen Island, the southwest area of Xiamen Island has the Gulangyu ferry terminal, South Putuo mountain, Xiamen University, and other attractive attractions, which also have a high probability of visiting. What is more, the western area outside of the Xiamen Island, the northern area outside of the Xiamen Island also has a high probability of visiting. However, the northern part of the study area has a relatively low probability of access, due to its relatively underdeveloped economy.

Figure 12 shows the distribution of inflow and outflow ratios of each hexagonal grid divided by period. It also can be seen from Figure 12 that, in the periods with a low number of drop-offs in the surrounding area, the central business district of Xiamen Island has the higher number of drop-offs; this also shows that, during working hours in a day, as the destination of most trips, there is a higher probability of drop-off in the area.

Figure 12. Cont.

Figure 12. Hourly distribution of the inflow and outflow ratio (P_{i_num}) in each hexagonal grid. (a–l) rerepresent 12 periods spectively in a day.

3.2. The Test Result of The Model

In this part, we selected the test data from the remaining 20% of data in the study area and chose three hexagonal grids in each period of the day, applied the TI-STPM model to all the real trips inside the grid to calculate the possible travel distance, and predicted its possible visiting location to test the accuracy of the TI-STPM model. First, we built a buffer prototype based on the radius of the travel distance, which is calculated based on the pickup time and location. Then we extracted the grid ID with the highest visit probability in the buffer and compared the extracted grid ID with the grid where the drop-off point in the real itinerary is located. Finally, we got the test result of the accuracy of the proposed model. Figure 13 shows the actual itinerary of the selected verification data, and Table 3 shows the accuracy of the proposed model.

Figure 13. Pickup location (left) and the drop-off location (right) of 36 verification grids.

Table 3. Comparison of prediction accuracies at TI-STPM and real data.

No.	Time_Period	Hex_ID	Travel_Num	Correct_Num	Accuracy (%)
01	00~02	60	119	53	44.54
02		19	126	51	40.48
03		82	150	64	42.67
04	02~04	39	492	250	50.81
05		46	41	22	53.66
06		107	24	11	45.83
07	04~06	44	76	33	43.42
08		63	244	119	48.77
09		82	48	26	54.17
10	06~08	101	45	25	55.56
11		83	75	35	46.67
12		38	223	122	54.71
13	08~10	81	299	177	59.20
14		34	344	191	55.52
15		69	406	262	64.53
16	10~12	105	32	14	43.75
17		7	104	56	53.85
18		63	616	412	66.88
19	12~14	5	41	20	48.78
20		44	653	447	68.45
21		68	663	420	63.35
22	14~16	19	52	30	57.69
23		64	621	411	66.18
24		39	944	353	37.39
25	16~18	21	6	3	50.00
26		14	291	154	52.92
27		43	89	47	52.81
28	18~20	16	162	98	60.49
29		66	2555	1739	68.06
30		47	638	411	64.42
31	20~22	15	1155	770	66.67
32		8	23	11	47.83
33		57	30	16	53.33
34	22~24	23	11	6	54.55
35		25	19	10	52.63
36		61	801	485	60.55
Mean				54.19	

It can be seen from Table 3 that the accuracy rate of the calculation result is almost over 40%, the highest is 68.45%, and the average accuracy rate also reaches 54.2%.

Analyzing the results at different periods of the day, it can be seen that, in the period from 12:00 a.m. to 6:00 a.m. of the day, the forecast results are less than 50%, and this

may be because that urban bus is out of service during this period, so it is unreasonable to take urban bus lines as a reverse factor affecting taxi travel at this time. While the period between 6:00 a.m. and 10:00 p.m. is usually a period when urban human flows are more active, considering the competitive effect of urban buses on taxis during this period can help to improve the accuracy of model prediction to a certain extent; thus, in this period, the prediction accuracy rate is high.

In addition, it can be seen from the calculation results that, when there is a large number of taxi travels in the selected grid, the prediction accuracy rate of the model is also high. This result may be due to the fact that the grids with a large number of taxi orders are usually developed areas, and these areas usually have developed road networks and bus lines; these two factors can improve the accuracy of the model results, so there is a high model-prediction-accuracy rate in these areas.

4. Discussion and Conclusions

In previous studies, social-media data, land-use data, and vehicle-GPS data were combined with general probabilistic models to identify taxi destinations in the study of urban-population-flow prediction. We based our research on these research results; nonetheless, our research goal was to find areas where a combination of multi-source datasets and temporal and spatial patterns cannot be used to predict where taxis will visit. Specifically, the purpose of our research was to construct a probability model through the overall analysis of human flows in time and space, and use the probability model to predict the probability of visits to urban locations when taking a taxi for a trip.

The temporal and spatial structure of urban spatial human flows, using taxis as a mode of travel, usually shows a combination of reasonable spatial distribution and significant temporal laws. The rational interpretation of the spatial distribution should be based on the external constraints of the human flows in the urban space, such as spatial distance, road network, etc. The interpretation of the law of time should be based on the internal drive in urban activities, such as service demand time. On the other hand, the urban spatial human-flow pattern can reflect the hot spots of urban, and that can help predict urban human flows more reasonably.

Urban human flows can be predicted by the spatial-distribution characteristics and time-distribution law of daily traffic travel. Mining the inherent characteristics of traffic travel data can help us discover the complex relationship of human flows in time and space, especially after the detailed division of timescale and space scale; the fine-grained characteristics shown by it show the difference of urban human flows. In this research, we conducted fine-grained feature mining on urban traffic data (taxi travel data) and performed a feature analysis on travel distances and the number of trips in different periods to build a model for predicting access probability, which can help to improve the accuracy of urban human-flow forecasts.

This study used urban taxi travel data to extract the temporal and spatial characteristics of the human flows and then combined them with road networks and bus lines to construct a temporal and spatial probability model to predict location visits when taking taxis as a travel mode. Since most of the predictions of location visits are based on the types of activities that humans participate in and the types of urban land use, we tried to predict the location of visits by proposing a spatiotemporal probability model, regardless of the type of activity and land use. The model is based on the spatiotemporal clustering characteristics of human taxi activities analyzed by the k-means clustering method and the kernel density analyzing method, taking into account the number of trips, travel distance, travel-time distribution, urban-road-network density, and urban-bus-line density as model calculation factors to predict the possibility of location visits.

The results show that the proposed model can successfully predict where people are likely to visit when traveling by taxi, and the average prediction accuracy reached 54%. Although the prediction accuracy is not high, this method provides an idea for studying the prediction of visiting locations in cities, and the research results could help to take more

scientific and reasonable measures in urban management. For example, in areas with a high frequency of visits, public-service resources should be increased to ensure normal human flows. The accuracy of the model is limited by the particularity of the geographical location of the study area; for datasets with more general characteristics, this method may have higher prediction results. In addition, factors such as traffic control and road restrictions can be added to the calculation of the model to achieve higher prediction accuracy.

Author Contributions: Fu Ren proposed the original idea and conducted the organization of the content. Wenwen He carried out experiments and analysis of the results and wrote the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China (Project No. 42071448).

Data Availability Statement: Not applicable.

Acknowledgments: Thank the reviewers and editors for their constructive comments on this paper.

Conflicts of Interest: The authors declare no conflict interest.

References

1. Wang, Z.; Yue, Y.; He, B.; Nie, K.; Tu, W.; Du, Q.; Li, Q. A Bayesian spatio-temporal model to analyzing the stability of patterns of population distribution in an urban space using mobile phone data. *Int. J. Geogr. Inf. Sci.* **2020**, *35*, 1–19. [\[CrossRef\]](#)
2. Hasan, S.; Zhan, X.; Ukkusuri, S.V. Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, IL, USA, 11 August 2013; pp. 1–8.
3. Roth, C.; Kang, S.M.; Batty, M.; Barthélemy, M. Structure of urban movements: Polycentric activity and entangled hierarchical flows. *PLoS ONE* **2011**, *6*, e15923. [\[CrossRef\]](#)
4. Steiger, E.; Resch, B.; Zipf, A. Exploration of spatiotemporal and semantic clusters of Twitter data using unsupervised neural networks. *Int. J. Geogr. Inf. Sci.* **2016**, *30*, 1694–1716. [\[CrossRef\]](#)
5. Yang, Z.; Franz, M.L.; Zhu, S.; Mahmoudi, J.; Nasri, A.; Zhang, L. Analysis of Washington, DC taxi demand using GPS and land-use data. *J. Transp. Geogr.* **2018**, *66*, 35–44. [\[CrossRef\]](#)
6. Zhao, P.X.; Qin, K.; Zhou, Q.; Liu, C.K.; Chen, Y.X. Detecting hotspots from taxi trajectory data using spatial cluster analysis. In Proceedings of the ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Fairfax, VA, USA, 13–15 July 2015; Volume 2, pp. 131–135.
7. Keler, A.; Krisp, J.M.; Ding, L. Extracting commuter-specific destination hotspots from trip destination data—Comparing the boro taxi service with Citi Bike in NYC. *Geo-Spat. Inf. Sci.* **2020**, *23*, 141–152. [\[CrossRef\]](#)
8. Zhang, T.; Li, Y.; Yang, H.; Cui, C.; Li, J.; Qiao, Q. Identifying primary public transit corridors using multi-source big transit data. *Int. J. Geogr. Inf. Sci.* **2018**, *34*, 1137–1161. [\[CrossRef\]](#)
9. Jiang, Z.; Evans, M.; Oliver, D.; Shekhar, S. Identifying K Primary Corridors from urban bicycle GPS trajectories on a road network. *Inf. Syst.* **2016**, *57*, 142–159. [\[CrossRef\]](#)
10. Rahmani, M.; Koutsopoulos, H.N. Path inference from sparse floating car data for urban networks. *Transp. Res. Part C Emerg. Technol.* **2013**, *30*, 41–54. [\[CrossRef\]](#)
11. Zhang, S.; Tang, J.; Wang, H.; Wang, Y.; An, S. Revealing intra-urban travel patterns and service ranges from taxi trajectories. *J. Transp. Geogr.* **2017**, *61*, 72–86. [\[CrossRef\]](#)
12. Tang, J.; Zhang, S.; Zhang, W.; Liu, F.; Zhang, W.; Wang, Y. Statistical properties of urban mobility from location-based travel networks. *Phys. A Stat. Mech. Its Appl.* **2016**, *461*, 694–707. [\[CrossRef\]](#)
13. Castro, P.S.; Zhang, D.; Li, S. Urban traffic modelling and prediction using large scale taxi gps traces. In Proceedings of the Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Newcastle, UK, 18–22 June 2012; Volume 7319, pp. 57–72.
14. Tang, J.; Jiang, H.; Li, Z.; Li, M.; Liu, F.; Wang, Y. A Two-Layer Model for Taxi Customer Searching Behaviors Using GPS Trajectory Data. *IEEE Trans. Intell. Transp. Syst.* **2016**, *17*, 3318–3324. [\[CrossRef\]](#)
15. Wong, R.C.P.; Szeto, W.Y.; Wong, S.C. A cell-based logit-opportunity taxi customer-search model. *Transp. Res. Part C Emerg. Technol.* **2014**, *48*, 84–96. [\[CrossRef\]](#)
16. Liu, L.; Andris, C.; Ratti, C. Uncovering cabdrivers' behavior patterns from their digital traces. *Comput. Environ. Urban Syst.* **2010**, *34*, 541–548. [\[CrossRef\]](#)
17. Reades, J.; Calabrese, F.; Ratti, C. Eigenplaces: Analysing cities using the space—Time structure of the mobile phone network. *Environ. Plan. B Plan. Des.* **2009**, *36*, 824–836. [\[CrossRef\]](#)
18. Zhou, Y.; Fang, Z.; Thill, J.C.; Li, Q.; Li, Y. Functionally critical locations in an urban transportation network: Identification and space-time analysis using taxi trajectories. *Comput. Environ. Urban Syst.* **2015**, *52*, 34–47. [\[CrossRef\]](#)

19. Tang, J.; Hu, J.; Wang, Y.; Huang, H.; Wang, Y. Estimating hotspots using a Gaussian mixture model from large-scale taxi GPS trace data. *Transp. Saf. Environ.* **2019**, *1*, 145–153. [[CrossRef](#)]
20. Tang, J.; Liang, J.; Zhang, S.; Huang, H.; Liu, F. Inferring driving trajectories based on probabilistic model from large scale taxi GPS data. *Phys. A Stat. Mech. Its Appl.* **2018**, *506*, 566–577. [[CrossRef](#)]
21. Zheng, Z.; Zhou, S. Scaling laws of spatial visitation frequency: Applications for trip frequency prediction. *Comput. Environ. Urban Syst.* **2017**, *64*, 332–343. [[CrossRef](#)]
22. Yue, Y.; Zhuang, Y.; Li, Q.; Mao, Q. Mining time-dependent attractive areas and movement patterns from taxi trajectory data. In Proceedings of the 2009 17th International Conference on Geoinformatics, Geoinformatics, Fairfax, VA, USA, 12–14 August 2009.
23. Gong, L.; Liu, X.; Wu, L.; Liu, Y. Inferring trip purposes and uncovering travel patterns from taxi trajectory data. *Cartogr. Geogr. Inf. Sci.* **2016**, *43*, 103–114. [[CrossRef](#)]
24. Huang, X.; Ye, Y.; Guo, H.; Cai, Y.; Zhang, H.; Li, Y. DSKmeans: A new kmeans-type approach to discriminative subspace clustering. *Knowl. Based Syst.* **2014**, *70*, 293–300. [[CrossRef](#)]
25. Huang, J.Z.; Ng, M.K.; Rong, H.; Li, Z. Automated variable weighting in k-means type clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 657–668. [[CrossRef](#)] [[PubMed](#)]
26. Jiawei, H.M.; Kamber, P.J. *Data Mining: Concepts and Techniques*; Elsevier: Amsterdam, The Netherlands, 2011.
27. Thorndike, R.L. Who belongs in the family? *Psychometrika* **1953**, *18*, 267–276. [[CrossRef](#)]
28. Birch, C.P.D.; Oom, S.P.; Beecham, J.A. Rectangular and hexagonal grids used for observation, experiment and simulation in ecology. *Ecol. Modell.* **2007**, *206*, 347–359. [[CrossRef](#)]