Smart City – Taxi Passenger Aboard Location Recommendation

Abstract: With the fast development of society, people's life is becoming more and more convenient, and people's demand for travelling are higher and higher. Due to the convenience, taxi has become one of the most common transportation vehicle. When finding available taxis, it is common that passengers have to wait for a long period of time. However, simply increasing the amount of city taxis would only decrease the success rate of obtaining passengers and energy utility rate. By analyzing and mining the GPS tracking data of taxi, we can get an insight of taxi passenger's location distribution and predict the hot spot of passenger location at the same time. Codes are available on GitHub.

Introduction: In this project, we analyze GPS tracking data of more than 4000 taxis in Shanghai, China on 2th February, 2007. By using clustering method of DBSCAN to extract and analyze the hot spot of taxi passenger location, it is found the distribution of host spot location varies a lot in different time. On top of that, we build a taxi passenger location recommendation model based on time. By calculating the recommendation rate of different hot spot location and distance of taxi, we recommend the passenger aboard location for drivers, which would help to increase the profit of drivers, increase energy utility rate and alleviate city pollution problem at the same time.

1. Data analysis

In data preprocessing stage, we removed duplicated data, data not in the research area due to equipment of vehicle issue and convert time information into scale of hour. GPS tracking data of taxi is continuous in time., so the tracking data can reflect the change of passenger aboard. The 'State' variable in data means whether are passengers in the taxi when driving. 1 means there is passenger and 0 means no. According to the change of 'state' along time, we can determine the passenger aboard location, i.e., the 'State' changing from 0 to 1 is the location where passengers aboard.

2. Clustering of taxi aggregating location

DBSCAN is a clustering algorithm based on density distribution. We counted the number of actual passenger aboard location in different time and found that the travelling demand is highly correlated with the time. From 0 to 7 a.m., few people are travelling outside. Most people travelling from 8 a.m. to 18 p.m., the amount decreases after that and on 20 p.m., there's another small peak. Therefore, clustering based on different time is required. We visualized the clustering result of 9 a.m. and 18 p.m. from latitude 31.20~31.25 and longitude 121.5~121.55 in Shanghai, as the Figure 1. It can be observed that the taxi aboard location has big difference in time and space.

3 Passenger aboard location recommendation algorithm based on hot spot district

More centralized passenger aboard location means higher possibility to get passenger. But we can't simply recommend the host spot location to drivers. In this project, we propose a method to calculate recommendation rate to predict success rate of obtaining passenger as below:

(1) Let Sum_{point} denote number of passenger aboard location in a hot spot district. We calculate the central point of

each hot spot district as blow, X_i and Y_i denotes the longitude and latitude of passenger aboard location.

$$Center(lon, lat) = (\frac{1}{m}\sum_{i=1}^{m} X_i, \frac{1}{m}\sum_{i=1}^{m} Y_i)$$

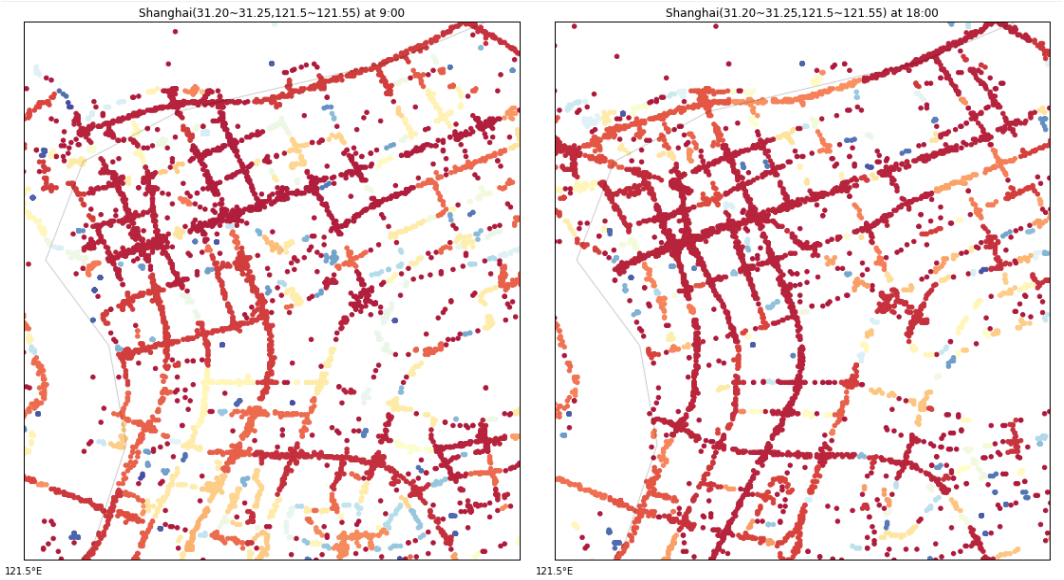

Figure 1

(2) Count number of taxis in each hot spot district as Sum_{taxi}

(3) Calculate the recommendation rate of each hot spot district as Value, defined as $\text{Value} = \text{Sum}_{\text{point}}/Sum_{taxi}$

Higher value means higher success rate of obtaining passengers in the hot spot district. The key point of the passenger aboard location recommendation system is that, finding the hot spot of taxi aggregation, and then recommend the location of higher Value and near distance to drivers.

In this project, we set the threshold of Value to be 0.1. So Value<0.1 means there are much more taxis then passengers in the district and success rate of obtaining passenger is very low. And Value>0.1 is the recommended passenger aboard location for driver. We visualized the passenger aboard location of 17 p.m. as Figure 2.

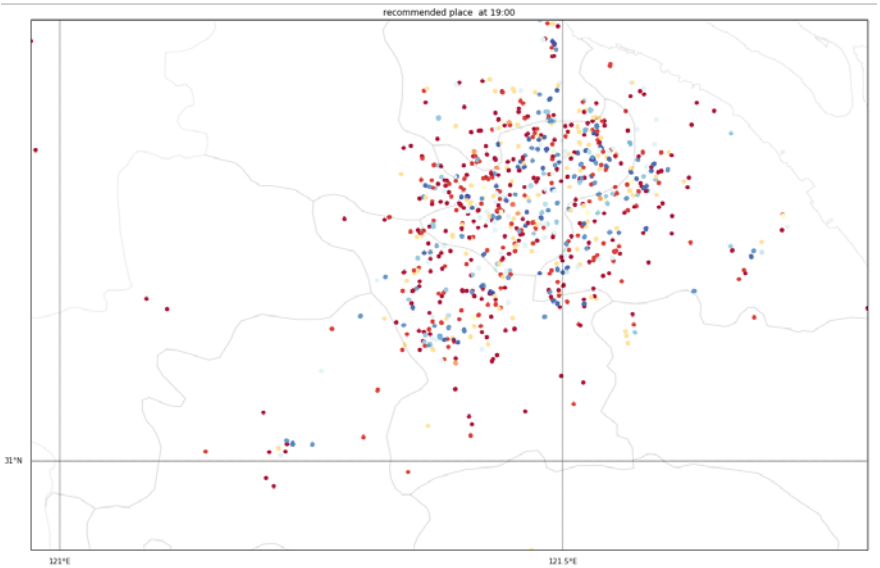

Figure 2