5/3/2015

Taxi Data Analysis

95-879 Mobile Intelligence and Business

Qifan Shi
HEINZ COLLEGE, CARNEGIE MELLON UNIVERSITY

Table of Contents

Problem Statement

This research targets on analyzing potential ways to increase taxi income. Basically, the research approaches this problem from three aspects: understand passenger's behavior, detect golden area, and give routing recommendations.

Understanding passenger's behavior: this is essentially performing trajectory visualization. This gives taxi driver basic information on demand of taxi trips by looking for trajectories patterns.

Detection of golden area: this relies on analysis of start and end of taxi trip. The golden area is defined as the area that either contains the start or end point of a taxi trip, or both. It is built upon the first aspect but a step further. Rather than visualization for understanding behavior, the area is detected from the quantitative method which building a social network of taxi trips. It gives more insight on areas that are heavily involved in taxi trips, which helps taxi driver on decision making about the best area he/she should go.

Routing recommendation: the last aspect depicts a vision of how to optimize taxi trip based on the social graph generated in detection of golden area. Rather than simply look if an area is the start or end of a trip, the recommendation requires an attention on the direction of trips. Given a trip request from place A to B, whether a taxi driver should go depends on if B area is the start of any trip. Such an analysis benefits taxi driver by minimized distance of driving without passengers.

This research mainly uses R as the tool for data analysis and modeling. Visualization is performed via both R and QGIS.

Data Understanding

The available datasets are two: one with taxi GPS data and another with income data of each taxi trip (**Exhibit 1**). This research mainly focus on the taxi GPS data which gives a picture of taxi trip trajectories. The GPS data size is approximately 33 GB collected from Shenzhen, a metropolitan city in southern China. It contains 603,660,071 rows of location records from 18,897 taxicabs during September 2009.

Below is a holistic view of taxi trajectory from taxicab with taxi number of B041C2 on September 1st 2009. The blue contour lines mark dense area of GPS location data, which depicts a rough outline on main driving area of the taxi on September 1st.

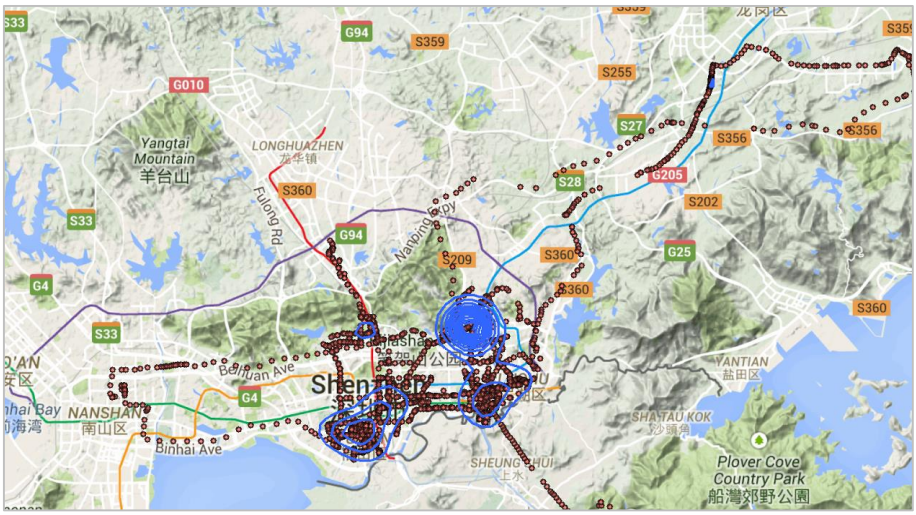


Figure 1 Holistic view of trajectory of taxi "B041C2" on 09/01/2009

Data Preprocessing

Subset: since the entire dataset contains more than 6 hundred million rows, using the entire data for analysis may significantly slow the retrieving and processing runtime in R. Also considering the limited course time and human resource, this research only uses a subset data from September 8th and 11th.

Occupation: the raw GPS data contains data both carrying passengers and not. This research only focuses on the data with passengers, free driving without passengers or personal-purpose trips are not within the scope of interest. The dataset is filtered by selecting "occupy" = 1.

Intersection: another preprocessing is to find the intersection of taxi trajectory and income. Initially this research aims at integration with income data even though it is not implemented within the 7-week time period. Therefore the GPS dataset is again filtered by selecting taxi numbers that both exist in income dataset and GPS dataset on the same day.

Taxi Type: in Shenzhen, there are three types of taxi colored as red, green and yellow. Different types of taxis has different driving zone and fare rate. Since red has no limitation on driving area and it takes the majority of taxis, this research further filter the data by concentrating on red taxi in Shenzhen, which is done by setting "taxi_type" = 1.

Task 1 - Understand Passenger's Behavior

Plotting all location data in map is a simple and efficient way to understand passenger's behavior. To simplify the operation, this research takes advantage of QGIS - an open-source geographical information system application - to visualize taxi trajectories. Rather than plotting on Google Map, the task uses the road shape file of Shenzhen. Below is the plotting of entire data from September 8th.

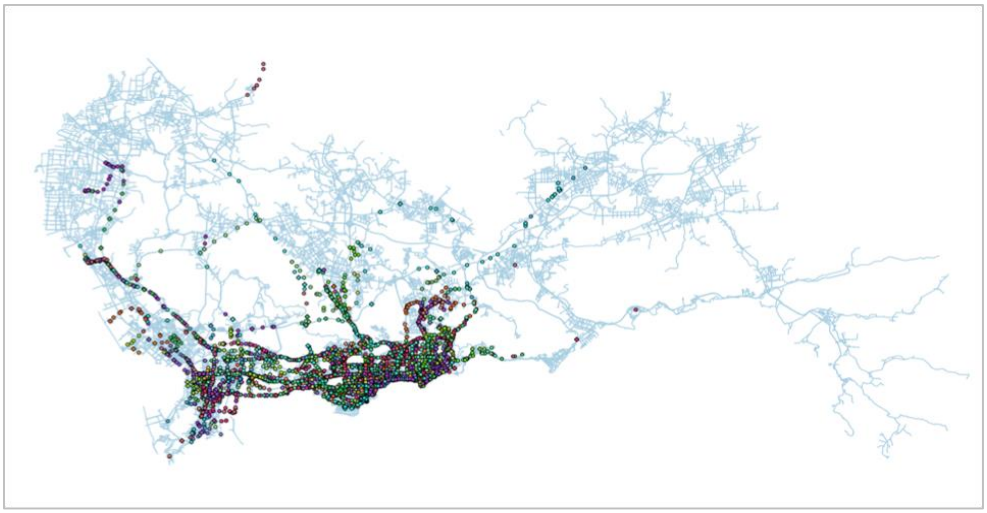


Figure 2 Taxi Trajectory on 09/08/2009

From the geo-image, it is easy to tell the distribution of taxi trajectories. The majority of taxis are active in west-southern part and a few scatter to north area. Eastern Shenzhen is not covered by any taxi trip during that day.

To give a more insightful observation, the task breaks down the entire-day view into hourly view. Also, for benchmarking purpose, the task introduces a comparison of hourly view between September 8th and 11th from 0 - 3 AM.

Two behavior can be detected from the breakdown of taxi trajectories (**Exhibit 2**) on September 8th. First, the taxi trips are in the lowest level from 3 - 6 AM, which reflect a low taxi demand at that time. Therefore driving a taxi 24 hours during that day is not a wise choice because the driver may not be able to pick up any passenger during low demand period. Also, if a taxi driver tries to be an "early bird", he should not go until 6 or 7 AM in the morning. Otherwise it will be a waste of time and gasoline.

Another interesting behavior reflects that people do not take taxi to and from airport 24/7. The hourly visualization shows two intervals: from 9 to 12 in the morning and from 6 - 11 in the evening. Other time people may prefer alternative ground transportation or just not go to airport.

To be more diverse, this research makes a comparison between September 8th and 11th data from 0 - 3 AM (**Exhibit 3**). The result is quite surprising. Both of the two days are work days - Tuesday of 8th and Friday of 11th, however, the taxi-trip distribution diverges a lot from each other. On 11th early morning, the taxi trips are much more than what on 8th. Therefore any taxi driver should not end their day of work early on September 10th if they desire to catch up potential income.

Task 2 - Detect Golden Area

In the detection of golden area, the research uses the GPS dataset from September 11th. The implementation is broken down into 3 steps:

- Map segmentation
- Trip extraction
- Social graph building

First, a clear definition of golden area is needed. In this research, the golden area is defined below:

> **Golden area:** the area that is heavily involved in taxi trips, it should be either as a start or an end of a trip, or both.

Map segmentation

This is a creation of area which forms a pool for selection of golden area. The entire geographical land of Shenzhen is enclosed with a smallest rectangle. Then a segmentation with 6 rows and 10 columns divides the entire map into 60 areas labels from 1 to 60. A graphical demonstration of map segmentation is below:

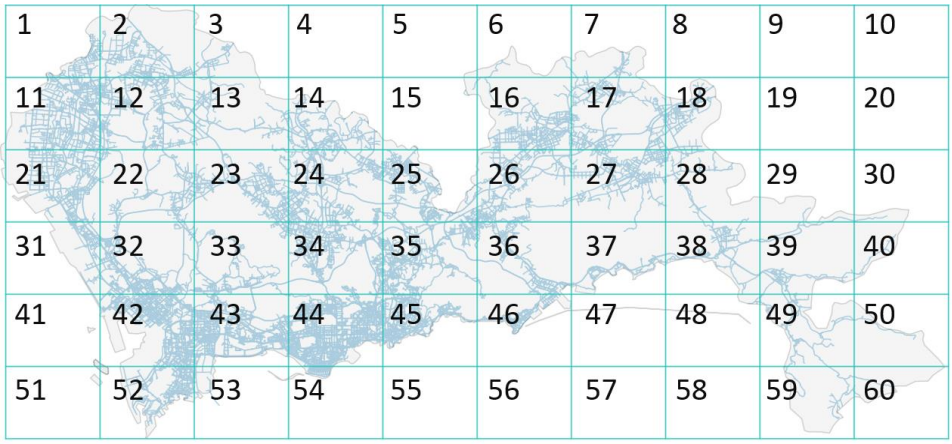


Figure 3 Map Segmentation of Shenzhen

Trip Extraction

A key challenge is to get a trip from the taxi GPS dataset. When detecting whether an area is a golden area or not, the only thing matters is if the area is the start or end, or both of a trip. There the entire trajectory of a trip is not necessary except the start location and end location.

The GPS data is uploaded in a relatively fixed interval, some taxis upload their location every 40 seconds, others go up to 50 seconds, and some even reach 80 seconds. However, in most cases, gaps between two trips are far more than 80 seconds. This research picks up a threshold[1] value of 300 seconds. Within each set of GPS data representing for a single trip, the first GPS location in the series is chosen as the start position and the last GPS location is chosen as end location of this trip. A new dataset with trips information is created after this step: each row stands for a trip with longitude and latitude for the start and end of the trip (**Exhibit 4 & 5**).

[1] The threshold is only used for splitting trips of the same taxi, the GPS data is first divided by taxi number. To be noticed this method works because the data is first filtered by occupation. So the GPS data does not continues every 40 , 50or 80 seconds.

Social graph building

The last step is to create a directed graph[2] with each node stands for a single area and edge stands for the number of trips between any two areas. The graph is implemented through a matrix with dimension of 60 × 60. The row number is defined as start location of a trip and column number is defined as end. For each of trips in the dataset generated in above step, a simple calculations is computed so that the start and end location are fallen into the corresponding areas. A mockup matrix is below:

Start \ End	1	2	3	4	5	6
1	0	**10**	0	0	**6**	0
2	0	0	0	0	0	0
3	0	0	0	**4**	0	0
4	0	0	0	**20**	0	0
5	**7**	0	**8**	**2**	0	0
6	0	0	0	0	0	0

Figure 4 Adjacent matrix of social graph

A better visualization of this matrix is to transform it back to a graph (**Exhibit 6**). A partial graph is showed below.

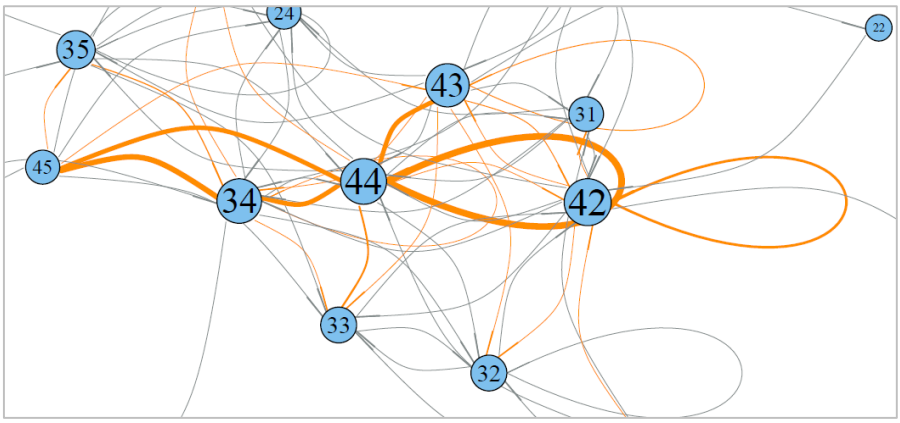


Figure 5 Partial social graph

In the graph displayed in figure 5, the size of the nodes implies the amount of trips start or end in this area. The higher the number of trips involved, the larger the size of the node. For edges, the color indicates the number of trips between connected two areas. Lower amount of trips colored as gray and higher amount of trips colored as orange. Among orange edges, the weight represents the orders, I which the strongest edges stands for the largest amount of trips. Golden areas are essentially the nodes with largest sizes the graph.

[2] Just for golden area an undirected graph is good enough. Here using a directed graph is due to the consideration for the third task.

From the graph, it is not difficult to tell that the area 44, 42, 43 and 34 are the 4 golden areas. This gives taxi drivers a suggestion on where they should go. No doubt on September 11th, they should go the above 4 areas because there are higher demand of catching a taxi in those areas.

To be noticed, even though 44, 42, 43 and 34 are all golden areas, but the composition of trips are different. Area 44, 42 and 43 contains many trips looping within themselves while the majority of trips in 34 connect to surrounding areas. Such a difference give a polymorphism of golden area, **star type**, **looping type** or a **mix type**. Star type basically refers to areas that connect to surrounding areas, such as 34. Looping type means areas that have trips travel within the same area, mix type simply means the combination of star and looping. All 44, 42 and 43 are mix type.

A more interesting follow-up of golden area is to perform an hourly golden area detection. This research selects a sample of 0 - 3 AM in September 11th for the hourly breakdown. However, the hourly golden area from 0 - 1 AM, 1 - 2 AM and 2 - 3 AM are all the same with 34, 42, 43 and 44, which does not show a shift in golden area. This may result from the period from 0 - 3 AM is essentially the same in late night. If the hourly analysis performs to the entire day, a shift of golden area may occur.

Even though the result in this task does not show a transition in golden area, such an approach still gives taxi driver help on better obtain taxi income. With the hourly golden area, the taxi driver can directly drive to the best area according to their own working hour. Assume a taxi driver starts to work at 8 AM, the best choice may be area 44. While another taxi driver starts to work at 3 PM may find area 34 is the best area.

Task 3 - Recommend Routing

The routing recommendation is a very useful decision making model for taxi driver. This research is not able to complete this task but it provides a vision on how such a recommendation system works. Below is the graph transformed from a part of the social graph on September 11th from 2 - 3 AM.

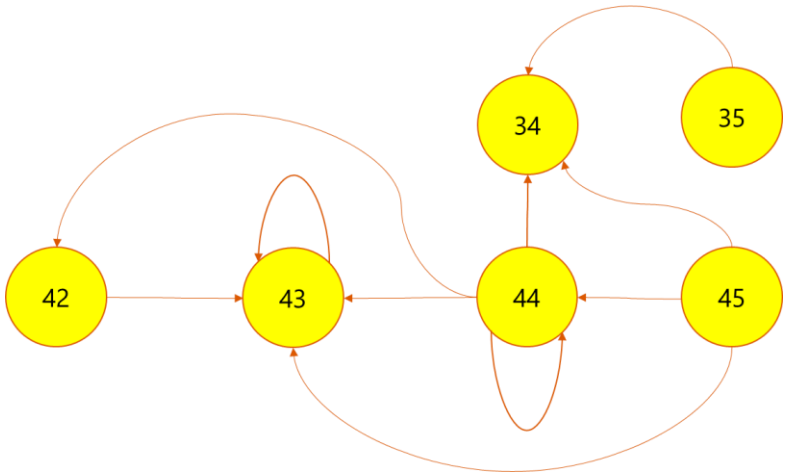


Figure 6 Partial social graph at 2 - 3 AM Sept. 11

Assuming a taxi is now in area 35, now there are two customers coming. One requests a trip to 34 and another request a trip to 44, which trip should the taxi go? Based on the graph show in figure 6, even though the trips to 34 and 44 give exactly the same fare, the taxi should go to 44 instead of 34 because

there is no “out” trip from 34. It indicates if the taxi goes from 35 to 34, it will hardly to pick up a passenger from 34 to itself of other areas. In order to get the next trip, the taxi may have to drive to the neighbor area 44 without a customer, which makes the taxi drive lose potential income. However, if the taxi goes to 44 where there are plenty of trips looping within 44 or connecting to other areas, the taxi may seamlessly get the next trip.

With enough historical information, the model can give proper recommendations to driver so that the driver can catch up with maximum fare.

Analysis and Conclusion

Overall, by performing task 1 and task 2, along with a shallow dive of task 3, this research is able to give taxi drivers a basic understanding of people’s travel behavior and guidance on the best area to go in an hourly basis. Task 3 illustrates a promising direction of research in the future, especially for taxi drivers collaborate with mobile application platform such as Uber or Didi.

It also has to be noticed that this research aims at providing an approach to increase taxi income by leveraging historical taxi trajectory data and income data. For the sake of efficiency and simplification, this research only uses a small subset from the raw dataset, which limits the accuracy of results as well as not feasible for any prediction purposes.

Appendix

Exhibit 1 Taxi dataset

Taxi GPS data

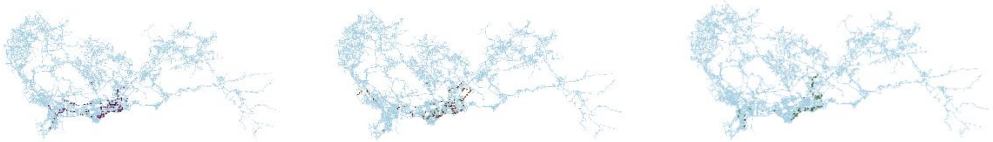
Name	Sample Value	Description
taxi_no	B041C2	The car plate number, which uniquely identifies of each taxi car
date	2009-09-01	The date from which the location data is recorded
time	00:00:23	The specific time when the location coordinate is recorded
lon	114.107640	The longitude of location
lat	22.570170	The latitude of location
speed	16	The current driving speed (km/h) when the location date is recorded
angle	15	The angle between driving direction and north
occupy	0	The state if the taxi is carrying passengers when recorded. 0 means no passengers, 1 means carrying passengers
unknown	0	The meaning of this column is unclear

Taxi trip income data

Name	Sample Value	Description
taxi_no	B041C2	The car plate number, which uniquely identifies of each taxi car,
begin_date	9/1/2009	The date when the taxi trip starts
begin_time	00:10:34	The time when the taxi trip starts
end_date	9/1/2009	The date when the taxi trip ends
end_time	00:15:02	The time when taxi trip ends
end_price	3.12	The fare rate (CNY/km) in the trip recorded
task_distance	0.779000	The trajectory distance (km) of the trip recorded
task_amount	16.100000	The total taxi fare in the trip recorded
free_distance	3.977000	The driving distance without carrying a passenger before the trip recorded
company_id	7758	The unique identifier which the taxi belongs to
taxi_type	1	The type of taxi. 1 stands for red taxi, 2 stands for green taxi, 3 stands for yellow taxi

Exhibit 2 Hourly view of taxi trajectory on September 8th 2009

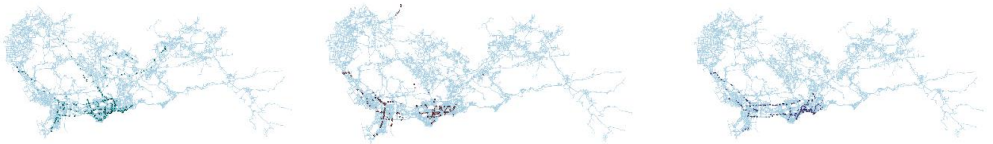
0 – 3 AM



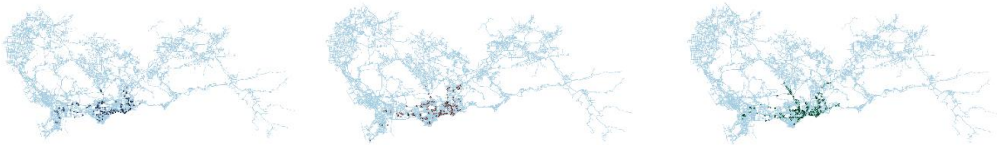
3 – 6 AM

6 – 9 AM

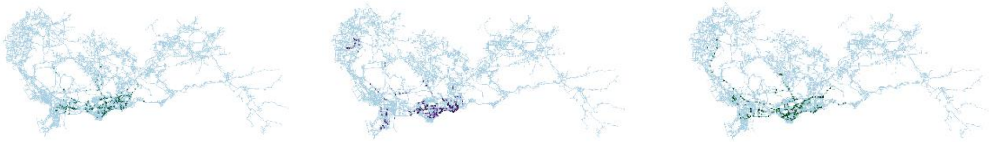
9 – 12 PM



12 – 3 PM



3 – 6 PM



6 – 9 PM

9 – 0 AM

Exhibit 3 Comparison of hourly taxi trip between Sept. 8th ad Sept. 11th

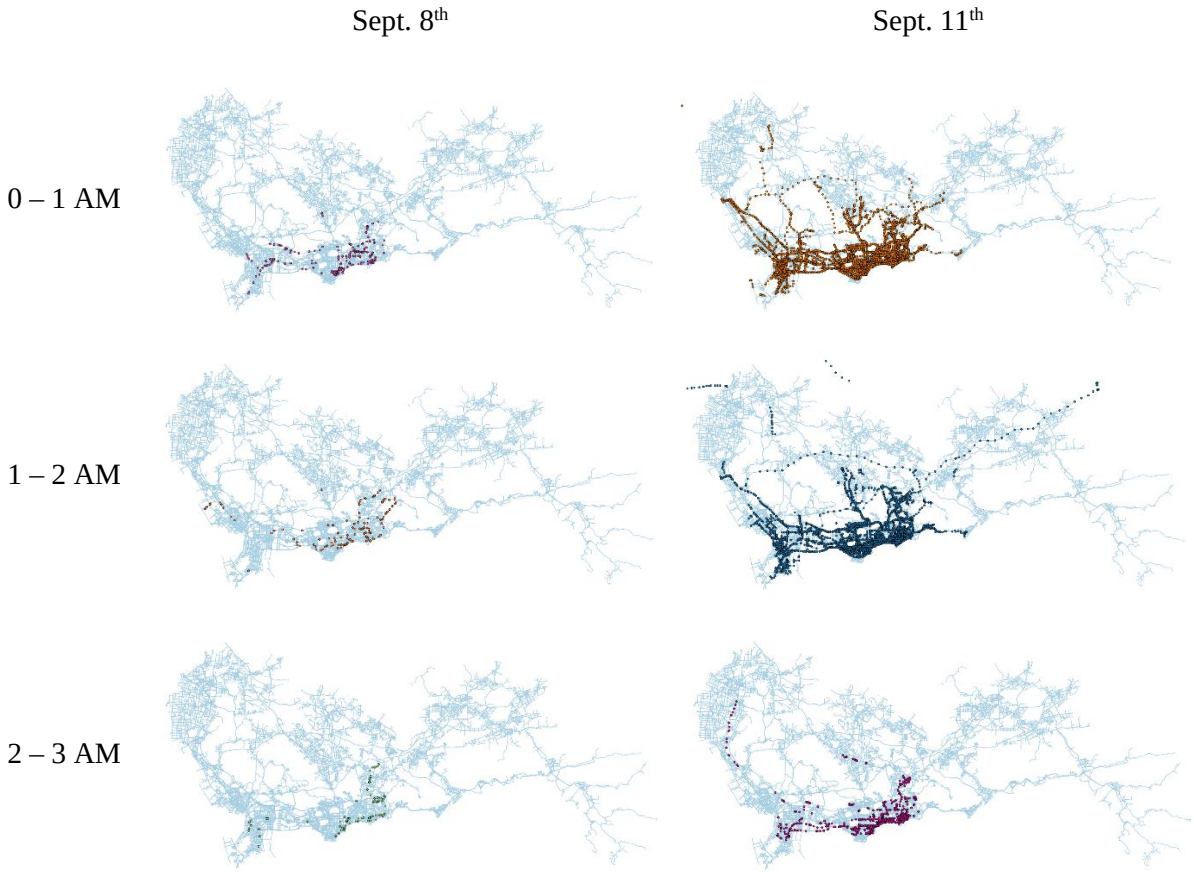


Exhibit 4 The whole view of taxi trip dataset

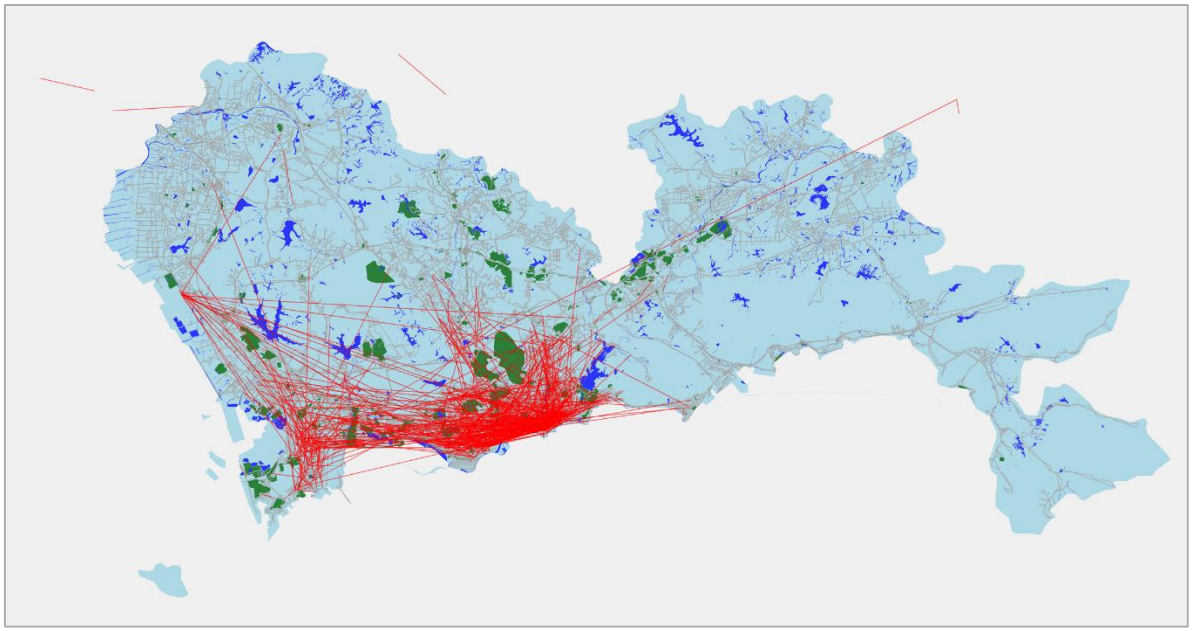


Exhibit 5 The trip dataset

Trip#	Start Longitude	Start Latitude	Destination Longitude	Destination Latitude
1	114.0492	22.52568	114.0371	22.52340
2	113.8307	22.65009	113.9186	22.52726
3	114.0991	22.57256	114.1011	22.60015

Exhibit 6 Social graph of taxi trip in September 11th