Recommendation for Pick-up Points and Hot Zone Based on New York Taxi Data

Ketong Xie

With the rapid development of wireless communication technology and intelligent mobile terminal, collecting the track record of moving objects becomes simple and fast. Due to the fact that the New York taxi trajectory data has not only spatial attributes, but also contains a time attribute, it becomes the research subjects and advanced application for spatial temporal data mining. Also, they can provide users location based services as well as the improvement for the urban planning and smart traffic. Based on the characteristics of taxi trajectory data being widely distributed and a large amount of data, it becomes an important research and application object of spatial temporal data mining. Providing taxi drivers driving recommendation for pick-up points and hot zone in New York not only helps to increase the economic income of the relevant personnel, but also reduces the burden of traffic caused by no-load taxi cruises, as well as reduces the unnecessary vehicle emissions for environmental protection.

However, the hot zone of passengers pick-up points will change with time, and at the same time, the city's available points will affect the passengers' travel pattern. How to accurately find the concentration of passengers in different periods of time is the necessary condition to the recommend services for taxi drivers. Firstly, this report will analysis the New York Taxi trajectory data to get the results of how people commute in the city. As mentioned before, the taxi trajectory data is big data so that the report also utilize the Spark on ROGER to calculate the results. After that, the report will use the methods of spatial autocorrelation to come up with the hot zone recommendation for New York and provide suggestion for city administration and taxi drivers in terms of improvement of the taxi services in these areas.

Data and Methodology

In this report, the New York taxi data containing over 14 million taxi records during January, 2013, is used to conduct the analysis. It is provided by the course of GEOG 479 in the Department of Geology. The taxi records of the data have detailed description of the taxi communication, including medallion, hack license, vendor id, rate code, store and fwd flag, pickup datetime, dropoff datetime, passenger count, trip time, trip distance, pickup longitude, pickup latitude, dropoff longitude, dropoff latitude, payment type, fare amount, surcharge, mta tax, tip amount, tolls amount, total amount. In this report, the descriptions of passenger count and pick-up location are mainly used to demonstrate the ridership in different traffic analysis zones with the help of the shapefile of the TAZ of New York. In future analysis, more research on the description of location and time periods can be conducted to analyze the change of taxi hot zones in different location and time periods. However, in this report, all the data will be considered to provide more general suggestion for taxi drivers and city administrators.

This report will mail use the methods of Spark and spatial autocorrelation. Spark is an open-source cluster computing framework originally developed in the AMPLabat UC Berkeley. The fundamental programming abstraction is Resilient Distributed Datasets (RDD), which is a logical collection of data partitioned across machines. It can run programs up to 100x faster Hadoop MapReduce in memory, or 10x faster on disk, which is quite helpful here to analyze the urban big data of taxi trajectories. Also, it brings ease of Use and write applications quickly in Java, Scala, Python, R. Spark runs on Hadoop, Mesos, standalone, or in the cloud. It can access diverse data sources including HDFS, Cassandra, HBase, and S3.

Besides, the report also uses the methods of spatial autocorrelation. Spatial autocorrelation definition measures how much close objects are in comparison with other close objects. Moran’s

I can be classified as: positive, negative and no spatial auto-correlation. Positive spatial autocorrelation is when similar values cluster together in a map. On the other hand, Negative spatial autocorrelation is when dissimilar values cluster together in a map. In this case, after the taxi trajectory data is analyzed on Spark, the results of total ridership in different TAZ will be father analyzed by the methods of spatial autocorrelation with help of R. As a result, the deliverables include maps (thematic, cluster maps, etc.), descriptive statistics, plots and graphics (histograms, QQ plots, etc.) and global and local spatial autocorrelation measures.

Analysis of Results

The results for the analysis on Spark is shown in Image 1. In the image, there are more ridership in several spots in the city. They are gathered in the western, central and southeast areas, which provide good incentives for taxi drivers to do business in these location. On the other hand, city administrators should also care for the traffic situation in these places because the high ridership could add to the pressure on the local traffics. In order to solve the problems, they can consider providing more public transportation or more capacity of the roads in these areas.

The results for the analysis on R is shown in Image 2. According to the Moran I test under randomization, the Moran I statistic is 0.9579998945, which is significantly close to 1, meaning that the ridership shows significant cluster characteristics. Moreover, there are 166 zones show such characteristics among 2246 zones in total. Based on Image 2 showing the results for LISA test in spatial autocorrelation, the 166 high-high clusters areas are mapped in the city. It shows that the clusters are mainly located in the western area. It is noted that there are also two other clusters in the central and southeast areas in the city, which have a large amount of scale compared to the western clusters.

Limitation and risks

Even though the results show significant characteristic in the report, there are still several limitation and risks in the process of the analysis. The report could also look into data in different location and time period in order to compare the change of the clusters. For example, in the analysis on Spark, the script can add some filters for drop off time such as the time in weekends, the time in morning and evening when people commute between work and families and the time in noon when there might be low needs for taxi drivers. By comparing the difference in these time periods, more insights for the taxi drivers and city administrators can be found to provide more helpful suggestion in terms of regulation and stimulus for the taxi industry.

On the other hand, there is another limitation and risk occurred in the analysis of spatial correlation. The report failed to look deeper into the results for spatial correlation in terms of relative factors. For example, the report could build a spatial correlation model to tell more details about the deciding factors for pick-up points and hot zone distribution. In order to fulfill that, the report could look for more demographic data such as employment, poverty, income level and total population in different TAZs. By doing so, a more detailed analysis of the distribution of pick-up points and hot zones will be conducted and reveals more information to taxi drivers and city administrators such as finding potential hot zones in the future.

Conclusion

The report used the methods of Spark and spatial autocorrelation to come up with the hot zone recommendation for New York and provide suggestion for city administration and taxi drivers in terms of improvement of the taxi services in these areas. The results for the analysis on Spark show that there is more ridership in the western, central and southeast areas, while the results for the analysis on R show that these areas show significant cluster characteristic. However, there are also several limitation and risks worth looking into in the future study, such as more filters on time periods and model building in spatial autocorrelation.

Image 1 New York Taxi Ridership

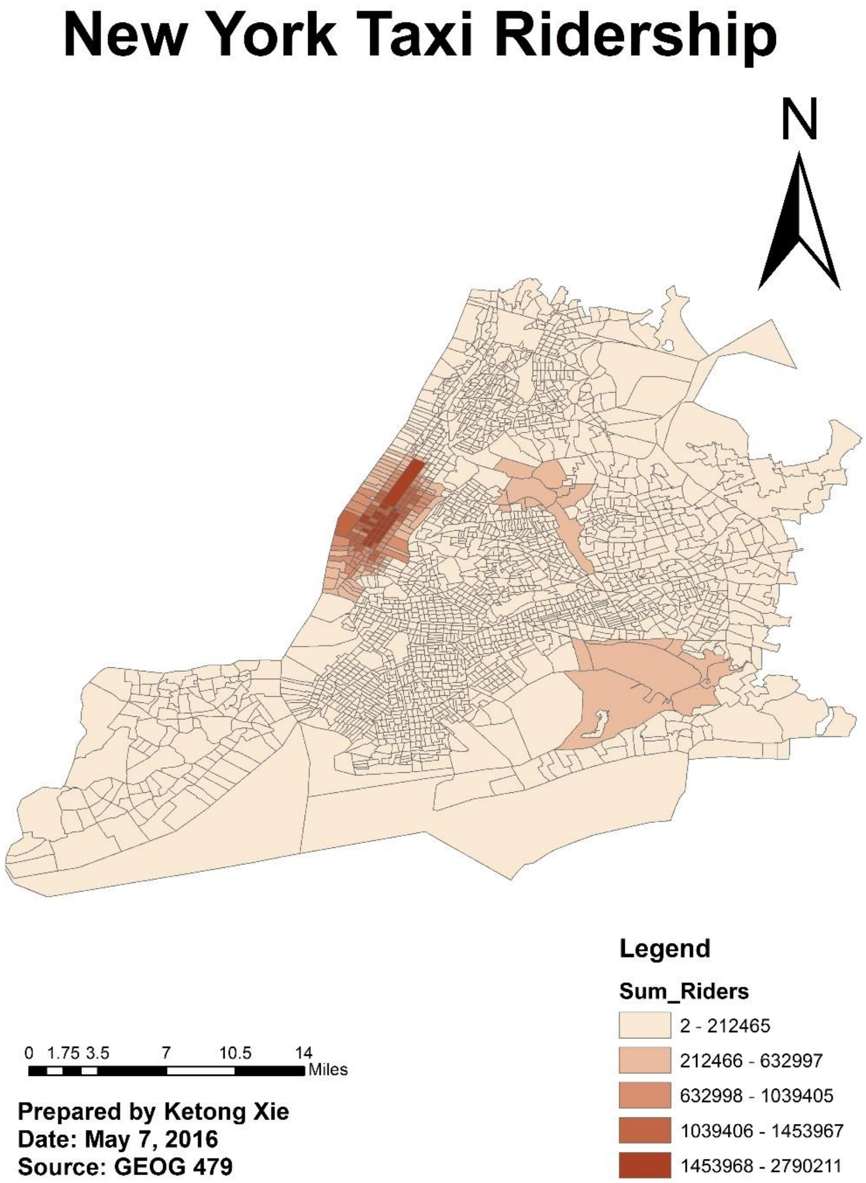
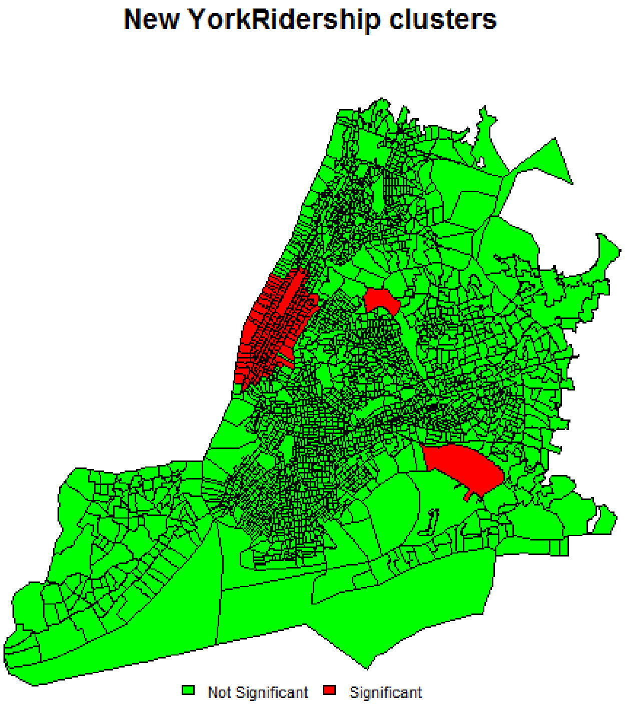


Image 2 New York Ridership Clusters


New YorkRidership clusters
🟩 Not Significant
🟥 Significant